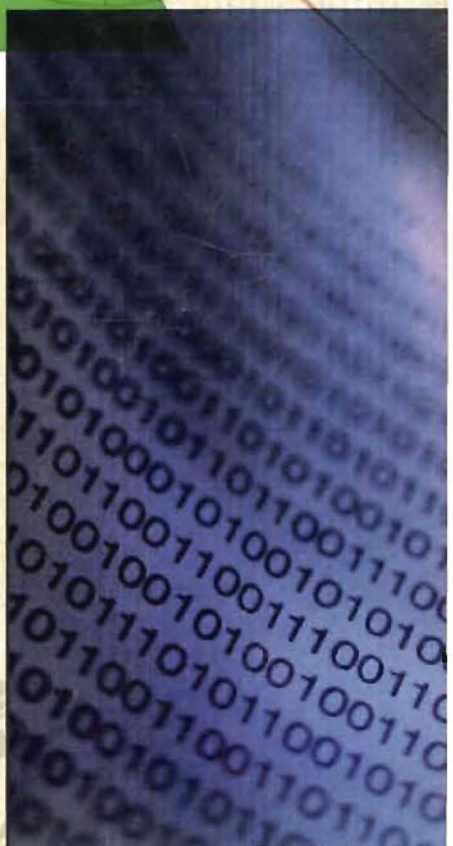
COLLOQUE FRANCOPHONE

enquêtes et systèmes d'information

LA STATISTIQUE EN ENTREPRISE



ABIDJAN/COTE D'IVOIRE
30 AVRIL 1999

CNSCA

IRD
Institut de recherche pour le développement

AFRISTAT

Association Internationale des Statisticiens

Société Française de Statistique

AFRISTAT - AISE - ENSEA - IRD - SFdS

dans le cadre du
Colloque francophone sur les

"Enquêtes et Systèmes d'Information"

Journée d'étude sur

LA STATISTIQUE EN ENTREPRISE

le 30 avril 1999

à
l'École Nationale Supérieure
de Statistique et d'Économie Appliquée
Abidjan, Côte d'Ivoire

Comité scientifique

Président :	
Lamine Diop	AFRISTAT - Mali
Xavier Charoy	INSEE - France
Jean-Jacques Droesbeke	ULB - Belgique
Anne-Marie Dussaix	ESSEC - France
Jean-Michel Gautier	HEC - France
Christian Gourieroux	CREST -France
Koffi N'Guessan	ENSEA - Côte d'Ivoire
Aka Kouame	IFORD - Cameroun
Ludovic Lebart	CNRS / ENST - France
Christophe Lefranc	INSEE - France
N'Cho Sombo	INS - Côte d'Ivoire
N'Diaye	Collège statistique-Sénégal
Adalbert Nshimyumuremyi	ENSEA-Côte d'Ivoire
Idrissa Ouattara	INS - Côte d'Ivoire
Marie Piron	IRD (ex-Orstom) - Côte d'Ivoire
Benoît Riandey	INED / AISE - France
Benjamin Zanou	ENSEA - Côte d'Ivoire

Sociétés ou instituts organisateurs

AFRISTAT - Observatoire Economique et Statistique d'Afrique Subsaharienne

AISE - Association Internationale des Statisticiens d'Enquêtes

ENSEA - Ecole Nationale Supérieure de Statistique et d'Economie Appliquée

IRD - Institut de Recherche pour le Développement (ex Orstom)

SFdS - Société Française de Statistique

Comité d'organisation

Président	
Koffi N'Guessan	ENSEA - Côte d'Ivoire
Secrétaire général	
Marie Piron	IRD - Côte d'Ivoire
Trésorière	
Sacy Nadaradjane	IRD - Côte d'Ivoire
Nicolas Reuge	ENSEA - Côte d'Ivoire
Thérèse Djabaté	IRD - Côte d'Ivoire
Michel Pépin	ENSEA - Côte d'Ivoire

Nos remerciements à :

Christelle Soumahoro, Patrick Bitty et Nadia Klemet (ICADJI)
pour leur participation à l'organisation de ce colloque
et toutes les personnes qui, de près ou de loin, nous ont aimablement soutenus.

AVANT PROPOS

Par les instituts organisateurs

Ecole Nationale Supérieure de Statistique et d'Economie Appliquée

Du 27 au 30 Avril 1999 se tiendra à l'Ecole Nationale Supérieure de Statistiques et d'Economie Appliquée (ENSEA) le colloque sur « *Enquêtes & Systèmes d'Information* ».

Ce colloque qui est une initiative conjointe de l'ENSEA et de l'Institut de Recherche pour le Développement (IRD ex ORSTOM) est organisé en collaboration avec l'Observatoire Economique et Statistique d'Afrique Sub-saharienne (AFRISTAT), l'Association Internationale des Statisticiens d'Enquêtes (AISE), et la Société Française de Statistique (SFDS).

Les travaux sont envisagés sur quatre (4) journées. D'une manière générale, ce colloque vise à (re)positionner les méthodologies d'enquête dans les pays en développement, et à soutenir la volonté d'améliorer le fonctionnement des systèmes statistiques en faisant découvrir de nouvelles approches de collectes des données ainsi que des méthodologies récentes. Une journée Entreprise est organisée afin de renforcer le contact entre la Statistique et l'Entreprise.

Au delà des rencontres et des relations qui seront nouées, il s'agit d'entreprendre une réflexion collective sur les systèmes d'information et leur impact sur la production des données, de l'information ainsi que leur gestion. De nombreuses personnes et des Entreprises ont réagi à l'appel à contribution diffusé à travers une plaquette élaborée à cet effet.

Les communications jugées pertinentes par rapport au thème de ce colloque ont été retenues par le Comité Scientifique. Je voudrais féliciter leurs auteurs et leur exprimer notre plaisir de pouvoir les accueillir à Abidjan.

Pour l'organisation du colloque, des organismes sollicités ont répondu favorablement en apportant une contribution financière plus ou moins importante : la Coopération Française, l'Institut de Recherche pour le Développement (IRD), AFRISTAT, la Banque Mondiale, EUROSTAT, l'AISE et la Société Ivoirienne de Raffinage (SIR).

Je voudrais aux noms du Comité d'Organisation et de la Direction de l'ENSEA, leur exprimer mes sincères remerciements pour cet appui financier qui permet de bénéficier et de partager l'expérience de personnes de notoriété reconnue sur ces questions.

J'adresse également mes vives félicitations aux membres du Comité d'Organisation et à toutes les personnes qui ont aidé à la tenue de ce colloque.

Enfin, je souhaite qu'à travers les échanges, des recommandations réalistes se fassent afin que leur mise en œuvre puissent permettre une amélioration significative de la production des données en Afrique.

Koffi N'GUESSAN
Directeur de l'ENSEA

Observatoire Economique et Statistique d'Afrique Subsaharienne

Le Colloque international sur " enquêtes et systèmes d'information " se tient à un moment où les systèmes statistiques nationaux (SSN) des pays francophones d'Afrique subsaharienne sont confrontés à une demande pressante tant de la part des décideurs aux niveaux national et régional que des bailleurs de fonds. Dans beaucoup de pays les moyens mis à disposition par les gouvernements pour faire face à une telle demande augmentent très peu s'ils ne stagnent pas.

Toutefois les nouvelles technologies de l'information qui deviennent de plus en plus accessibles et performantes constituent pour les statisticiens de ces pays une excellente opportunité notamment pour la collecte et la mobilisation des informations ainsi que leur diffusion.

Le colloque d'Abidjan est l'occasion de faire le point sur les expériences récentes dans le développement des systèmes d'information, dans l'amélioration de la qualité des données, du traitement et de la diffusion de l'information au service des décideurs.

En décidant de participer activement à l'organisation du colloque, AFRISTAT entend jouer pleinement son rôle dans l'amélioration et l'utilisation de l'information économique et sociale en Afrique subsaharienne.

Lamine DIOP
Directeur Général d'AFRISTAT

Institut de Recherche pour le Développement (ex Orstom)

Depuis 1985, l'Ecole Nationale Supérieure de Statistique et d'Economie Appliquée (ENSEA) d'Abidjan et l'Institut de Recherche pour le Développement (IRD) développent un partenariat qui repose sur des actions d'enseignement, de formation à la recherche, sur la réalisation de projets de recherche, et l'organisation de réunions scientifiques. En collaboration avec l'observatoire économique et statistique d'Afrique sub-saharienne (AFRISTAT), l'Association Internationale des Statisticiens d'Enquêtes (AISE) et la Société Française de Statistique (SFdS), l'ENSEA et l'IRD organisent un colloque sur un thème large et essentiel "Enquêtes et Systèmes d'Information".

Ce colloque se propose de réunir un certain nombre d'experts et d'expériences susceptibles d'éclairer la problématique particulière des systèmes d'information dans les pays en développement. En effet, quels que soient les domaines d'études les systèmes d'information ne cessent de se développer et de se complexifier. Mais améliorent-ils la qualité de l'information ? Permettent-ils d'appréhender et de traiter "autrement" une information statistique ? Ne faut-il pas repenser les méthodes d'enquête en fonction de la réalité économique et sociale de l'Afrique ?

Répondant aux objectifs de l'IRD, la vocation de ce colloque est à la fois globale, stratégique et internationale. Elle est globale par l'implication multiple de tous les acteurs intéressés et participants à ce colloque, chercheurs, enseignants, praticiens, entrepreneurs, bailleurs de fonds... Elle est stratégique par la réflexion fondamentale sur le rôle que les systèmes d'information sont amenés à jouer dans les dispositifs de gestion et de décision dans nos sociétés. Elle est internationale par la qualité des communications, du Sud et du Nord, sur un sujet sans frontière, même si une attention particulière est portée à l'Afrique.

Dans le cadre de ce colloque, la journée entreprise rendue possible par la participation de la SFdS, paraît du plus grand intérêt. Elle va permettre de communiquer avec les entreprises publiques et privées ivoiriennes pour qui l'information, sa gestion, son exploitation statistique sont essentielles. Cette démarche correspond à un des objectifs de l'IRD, à savoir : contribuer au développement par le transfert et l'utilisation rationnelle des connaissances scientifiques disponibles et mobilisables.

Remercions toutes les personnes et institutions qui ont aidé à la réalisation de ce colloque. Cependant, il n'est que justice de faire une mention toute particulière à Marie Piron de l'IRD et à Koffi N'Guessan de l'ENSEA pour leur initiative et pour l'ardeur déployée pour la concrétisation du colloque "*Enquêtes et Systèmes d'Information*" à Abidjan.

Alain MORLIERE
Représentant de l'IRD Côte d'Ivoire

l'Association Internationale des Statisticiens d'Enquêtes et la Société Française de Statistique

L'initiative de l'ENSEA et de l'IRD (ORSTOM) d'organiser ce *colloque Enquêtes et Systèmes d'information* a d'emblée reçu un accueil très favorable de l'AISE et de la SFdS car c'est en Afrique que nous devons discuter des méthodes d'enquêtes les plus adaptées à la réalité géographique, sociale et économique du continent. Par sa généralité, la théorie des sondages a une portée universelle, mais la pratique des sondages ne se lit pas dans un ouvrage de mathématiques, aussi indispensable qu'il soit. L'objet de l'enquête intéresse le statisticien d'enquête au moins autant que le jeu mathématique du hasard maîtrisé, et ce jeu même ne peut ignorer la diversité que les unités statistiques résidence principale, ménage, famille connaissent à travers le monde et même en côte d'Ivoire.

Cette richesse culturelle de la Côte d'Ivoire ajoute à la complexité des enquêtes. Elle oblige à repenser la signification concrète de notions aussi centrales que celle de base de sondage. Les enquêtes sur le secteur informel nous révèlent que la diversité économique n'est pas moindre que la diversité sociale. Déjà en France, la rareté des bases de sondage de ménages ou d'individus nous distingue des pays d'Europe du Nord dont les statisticiens disposent en permanence de registres non seulement d'entreprises, mais aussi de logements ou d'individus. Les enquêtes sur les sous-populations doivent le plus souvent se greffer sur un recensement récent. Cette contrainte n'est pas moins africaine que française, mais un principe de parcimonie et d'efficacité s'impose encore plus aux statisticiens africains, ce qui les amène à concevoir le plus globalement possible leur système d'information.

Là se situe l'intérêt scientifique propre à ce colloque. Lorsque l'information est rare et coûteuse, toute information préexistante doit être utilisée, toute information créée doit être insérée de façon systématique dans le capital des connaissances. C'est bien ce que signifie l'expression système d'information : les enquêtes doivent être articulées comme le dispositif MADIO de Madagascar nous le montre, comme AFRISTAT nous y invite. Les enquêtes ne peuvent être analysées indépendamment des autres informations disponibles: plus que l'introduction d'une information auxiliaire, c'est l'osmose des sources au sein de micro-simulations que j'évoque par là.

La Société Française de Statistique a coopéré activement à l'élaboration du programme scientifique du colloque. La Journée d'Etude sur la Statistique en Entreprise porte particulièrement la marque de son *Groupe Enquêtes, Modèles et Applications,* qui s'est beaucoup intéressé aux applications marketing des enquêtes. La SFdS a connu la chance d'accueillir bon nombre de collègues africains au premier colloque francophone sur les Sondages, puisqu'en juin 1997, une vingtaine de collègues d'outremer avaient pu participer aux petits cours du CEFIL à Libourne avant de se rendre au colloque de Rennes. Nous espérons bien des communications de collègues africains au second colloque francophone sur les sondages qui se tiendra à Bruxelles en juin 2000 à l'initiative de Jean-Jacques Droesbeke, mais aussi aux 33èmes Journées de Statistique qui se dérouleront en mai 2000 dans la ville prestigieuse de Fès.

Il me reste à remercier Monsieur Koffi N'Guessan et toute l'équipe de l'ENSEA de son chaleureux accueil, à féliciter Monsieur Lamine Diop et Marie Piron pour l'excellent programme scientifique, et à souhaiter à chacun un excellent colloque.

Benoît RIANDEY
Directeur exécutif de l'AISE
Président du Groupe SFdS "*Enquêtes, Modèles et Applications*"

SOMMAIRE

PREFACE

L'information est devenue très précieuse pour la gestion et le fonctionnement des Administrations ainsi que pour rendre les activités de recherche plus dynamiques.

Des méthodes robustes et des moyens technologiques puissants existent de nos jours pour collecter, traiter et diffuser cette information. La diversité des besoins et la recherche d'une meilleure qualité de l'information ont favorisé l'émergence de ces nouvelles technologies.

Dans ce document, sont rassemblés les résumés des communications qui alimenteront les discussions durant ce colloque sur «*Enquêtes et Systèmes d'Information*» dont les aspects statistiques et méthodologiques ont été davantage développés par les auteurs.

Cette orientation a été suscitée pour amener les responsables des systèmes statistiques africains et leurs partenaires à s'interroger encore une fois sur les entraves au fonctionnement de ces systèmes statistiques malgré la disponibilité des compétences nationales et les ressources matérielles.

Pourquoi accéder à la bonne information continue de demeurer un problème en Afrique ? Ce problème interpelle également les institutions de formation qui ont l'obligation d'adapter leurs programmes à l'évolution des nouvelles méthodes de collecte de l'information.

A cet effet, ce colloque qui est une opportunité pour informer un public de décideurs et d'acteurs dans ce domaine permettra d'engendrer des réflexions sur les choix à opérer dans la modification des programmes de formation, car les méthodes traditionnelles de collecte : recensement, enquête par sondage, etc... ont des limites que d'autres méthodes permettent aujourd'hui de corriger.

C'est aussi une occasion de rencontre avec les Entreprises qui ont des besoins insatisfaits en statistique nécessaire à leur politique de marketing et leur gestion.

Le continent africain accuse des insuffisances dans de nombreux domaines dont celui de l'information. Il est à espérer que les travaux du colloque et les publications qui en sortiront, puissent contribuer à rendre plus performants les systèmes statistiques.

Koffi N'GUESSAN
Directeur de l'ENSEA

PRESENTATION GENERALE DU COLLOQUE

Objectifs

Quels que soient les domaines d'études (santé, éducation, économie, démographie, finances, marketing, environnement, ...), les observatoires, les systèmes d'information géographiques, les systèmes d'alerte précoce, les systèmes d'information sur les marchés, ..., se développent de plus en plus.

Ces systèmes d'information, modifient-ils les pratiques d'enquêtes dans les pays en développement? Améliorent-ils la qualité des données et la qualité de l'information ? Permettent-ils d'appréhender, de construire et de traiter autrement une information statistique ?

Principaux thèmes

Autour de la problématique des systèmes d'information dans les pays en développement, tant dans les secteurs publics que privés, seront abordés les thèmes suivants :

- bases de sondage, techniques d'échantillonnage
- collecte et mobilisation de l'information
- nouvelles technologies d'enquête, de gestion de l'information, télédétection
- qualité des données, contrôles de terrain
- enquêtes longitudinales, panel
- observatoires, bases de données
- traitements des données, data mining, restitution de l'information
- marketing, étude de marché
- qualité de l'information, validation, méta-information
- diffusion, accès à l'information

Des conférenciers invités interviendront sur les méthodologies d'enquêtes et les systèmes d'information.

Public concerné

- Praticiens de la statistique, méthodologistes des enquêtes et des systèmes d'information.
- Chercheurs, ingénieurs, enseignants, formateurs, étudiants dans le traitement de l'information.
- Institutions publiques, organismes internationaux, entreprises privées, banques, assurances.

Programme sommaire du colloque "Enquêtes et Systèmes d'Information"

Ce colloque se propose de traiter, tant dans les secteurs publics que privés, des implications des systèmes d'information autour de 3 axes :

Mardi 27 avril:	*Développement des systèmes d'information*
Mercredi 28 avril:	*Evolution des pratiques et des modes de collecte d'informations*
Jeudi 29 avril:	*Nouveaux développements pour le traitement et la diffusion de l'information*

LA STATISTIQUE EN ENTREPRISE

Vendredi 30 avril 1999

La Statistique est généralement à la base du système d'information interne de l'entreprise. Elle est l'outil indispensable de tout décideur, mais reste néanmoins trop largement méconnue dans beaucoup d'entreprises de la sous région.

C'est fort de ce constat que la 4ème journée du colloque sur les "*Enquêtes et Systèmes d'Information*" est organisée autour du thème de la statistique en entreprises. Elle est centrée sur les multiples besoins de traitement statistique que l'on rencontre dans divers domaines : banque, assurance, finance, publicité, communication, grande distribution, télécommunication, transport etc. Cette journée s'adresse aux bureaux d'études, aux grandes entreprises africaines, plus particulièrement à leurs statisticiens et responsables d'information.

Les participants peuvent ainsi découvrir les développements les plus récents concernant leurs domaines d'intervention, et de ce fait ont la possibilité de mieux appréhender le rôle de la statistique eu sein de l'entreprise.

Nous espérons par ailleurs que cette journée, intégrée au colloque "enquêtes et Systèmes d'Information", contribuera à renforcer les liens entre chercheurs et praticiens de la place, permettant à chacun de travailler en tenant compte des attentes, des difficultés mais aussi des succès des autres. Cela concoure à la mise en place d'un réseau de compétences, permettant sans aucun doute le décloisonnement et le développement d'une recherche fondamentale trop souvent déconnectée de la réalité du praticien.

Nicolas REUGE
Enseignant à l'ENSEA

SYSTEMES DE DIFFUSION DE DONNEES FINANCIERES EN TEMPS REEL : Marchés financiers et données haute fréquence

Christian Gourieroux,
Professeur à l'Université Paris-Dauphine,
Directeur du Laboratoire Finance-Assurance
du Centre de Recherche en Economie et Statistique

INSEE - CREST - Laboratoire Finance-Assurance
15, boulevard Gabriel Pen, 92245 Malakoff Cédex - FRANCE
Tél : (33) 1 41 17 78 00 / Fax : (33) 1 41 17 76 66
E-mail : gouriero@ensae.fr

Le développement de systèmes électroniques d'appariement des ordres, d'enregistrements des échanges, de diffusion d'information statistique en temps réel a profondément modifié le fonctionnement des marches financiers. Le but de cet exposé est de présenter les algorithmes permettant de confronter les offres et les demandes, de décrire les lignes du carnet d'ordres, de comprendre les effets du système d'appariement sur les évolutions des prix et des volumes échanges. Ceci nous conduira à introduire divers résumés statistiques, dont certains sont diffusés en temps réel, citons: les fonctions de prix bid et ask, les mesures d'activité intra-journalière, les durées pondérées entre échanges, les volatilites... Finalement, sera discutée la modélisation jointe des divers risques liés à la liquidité et aux mouvements de prix.

Fonctionnement général

continu ou fixing

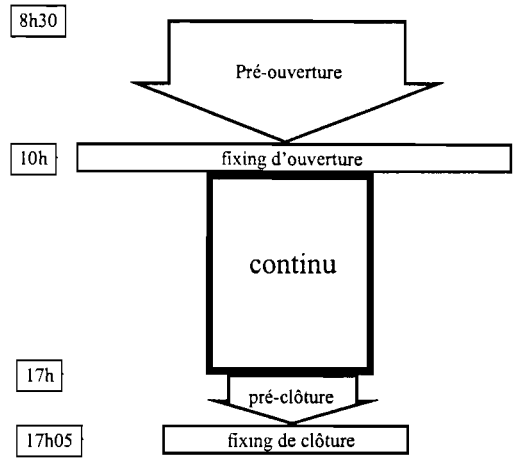
Les ordres arrivent en continu durant la séance et sont exécutés dès que possible.	Les ordres arrivent en continu durant la séance. Ils sont mis en attente pour être exécutés à une date prédéterminée.
⇔ les dates d'échange sont endogènes	⇔ les dates d'échange sont exogènes

Dirigé par les ordres ou par les prix

Pas d'intermédiaire. Le carnet d'ordres regroupe l'ensemble des ordres en attente, rangés par limite de prix et date d'arrivée.	Un intermédiaire (teneur de marché) affiche des prix d'achat (bid) de vente (ask).
Les investisseurs peuvent vendre ou acheter au bid ou ask du carnet.	Les investisseurs achètent ou vendent la quantité souhaitée à ces prix.

La bourse de Paris

Marché dirigé par les ordres sur lequel la cotation se fait selon une procédure mixte:

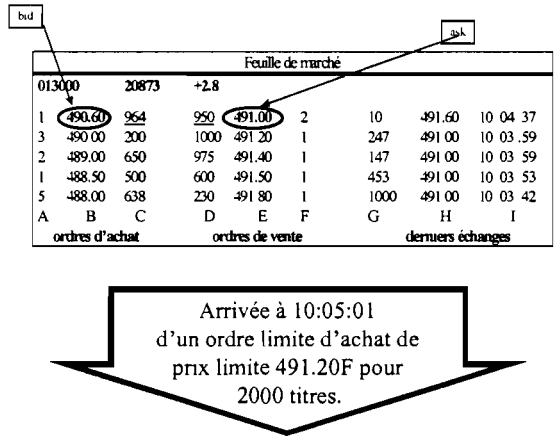


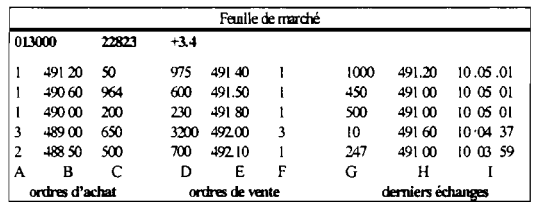
Exemples d'autres marchés

New York (ouverture au fixing, puis continu), Tokyo (ouvertures am et pm au fixing, puis continu).

Milan (ouverture en continu en général, puis fixing et retour en continu en fin de journée - Amihud-Mendelson-Murgia (1990))

Ordres et traitements des ordres





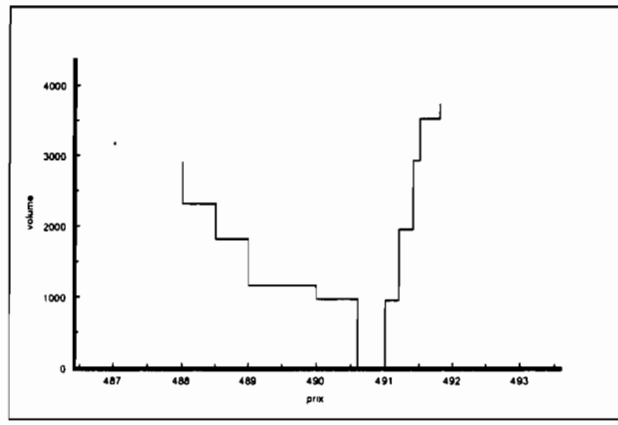
Feuille de marché									
013000		20873	+2.8						
1	490 60	964	950	491 00	2	10	491 60	10 04 37	
3	490 00	200	1000	491.20	1	247	491 00	10 03 59	
2	489 00	650	975	491 40	1	147	491 00	10 03 59	
1	488.50	500	600	491 50	1	453	491.00	10 03 53	
5	488 00	638	230	491.80	1	1000	491 00	10 03 42	
A	B	C	D	E	F	G	H	I	
ordres d'achat			ordres de vente			derniers échanges			

Arrivée à 10:05:01
d'un ordre d'achat au mieux
de 2000 titres.

Feuille de marché								
013000		22873	+3.6					
1	490 60	964	925	491.40	1	50	491 40	10·05·01
3	490 00	200	600	491 50	1	1000	491.20	10 05·01
2	489.00	650	230	491.80	1	450	491 00	10 05 01
1	488.50	500	3200	492.00	1	500	491.00	10.05 01
5	488.00	638	700	492.10	1	10	490.60	10 04 37
A	B	C	D	E	F	G	H	I
ordres d'achat			ordres de vente			derniers échanges		

Constitution de carnet d'ordres

Feuille de marché										
013000		20873		+2.8						
1	490 60	964	964	950	950	491 00	2	10	491 60	10 04 37
3	490 00	200	1164	1950	1000	491 20	1	247	491 00	10 03 59
2	489 00	650	1814	2925	975	491 40	1	147	491 00	10 03 59
1	488 50	500	2314	3525	600	491 50	1	453	491 00	10 03 53
5	488 00	638	2952	3755	230	491 80	1	1000	491 00	10 03 42
A	B	C		D	E	F	G	H	I	
ordres d'achat				ordres de vente			derniers échanges			



Effets des modes de négociation sur les échanges

Les systèmes de négociation diffèrent en termes de fréquence d'échange, de présence ou non d'intermédiaire, de prix d'échange, de transparence ... Quel est l'impact sur les échanges de ces différences ?

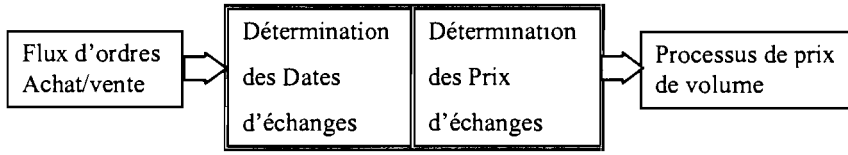
Deux effets : Effet direct et effet indirect

Effet direct + Effet indirect $\Rightarrow$ Effet global

Comment mesurer ces effets ?

But : Décrire et analyser l'effet direct du mode de négociation sur les caractéristiques de marché

Sur les marchés dirigés par les ordres, on s'intéresse à différentes procédures d'échange qui varient en termes de dates et de prix d'échange.



Dynamique du marché

- Le principe
  Ordres à prix limites $\begin{cases} \text{Achat : en } t,\ \alpha_t^* \text{ prix, } x_t^* \text{ volume,} \\ \text{vente : en } t,\ \beta_t^* \text{ prix, } y_t^* \text{ volume,} \end{cases}$

  unité monétaire [échelon de cotation]

  unité de volume [quotité]
- Cumul des ordres en attente

  $d_t(p) = x_t^* . 1\!\!1_{p \le \alpha_t^*}, s_t(p) = y_t^* . 1\!\!1_{p \ge \beta_t^*},$

  quantité maximale échangeable à ce prix

  $D_t^*(p) = D_t(p) + d_t(p)$

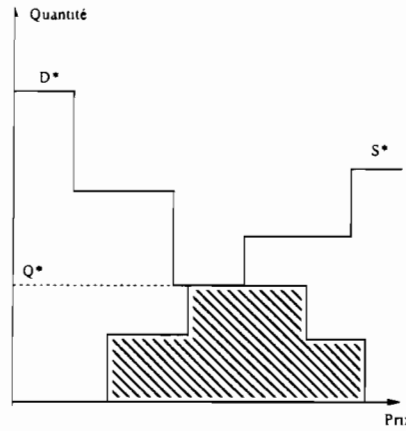
  $S_t^*(p) = S_t(p) + s_t(p)$

  $Q_t = \max_p \left[\min \left(S_t(p), D_t(p) \right) \right].$
- Y a t'il possibilités d'échanges ?

  $\Leftrightarrow$ Les courbes se croisent elles ?

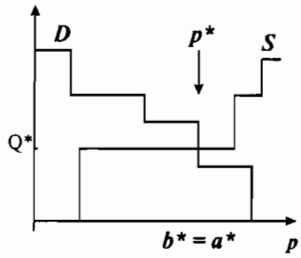
  $Q_t^* = 0$ pas d'échange à prix unique

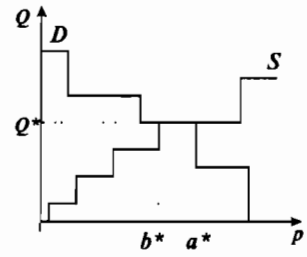
  $Q_t^* > 0$ possibilité d'échange



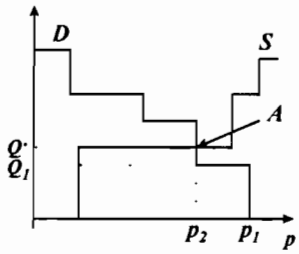
Prix d'échange

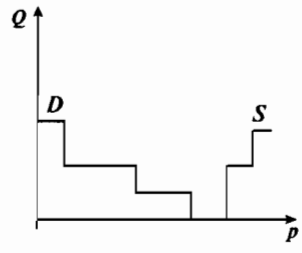
Prix unique d'exécution





Prix multiples d'exécution





$Q^* = Q_1 + Q_2$

Il y a troix règles à la Bourse de Paris pour déterminer le prix unique d'échange :

1) Maximisation du volume d'échange.
2) Minimisation du nombre d'investisseurs insatisfaits.
3) Ecart minimal entre prix d'échange et prix d'échange précédent.

Prix unique : par croisement des courbes de demande et offre on obtient l'ensemble des prix qui satisfont 1). *Ensemble* les prix sont discrets. Puis à l'aide des règles 2 et éventuellement 3, un prix unique d'échange est déterminé. Le plus proche de la notion de prix d'équilibre.

Prix multiples : utilisé en continu (à Paris).

Prix unique

Lorsque a*=b*, l'ensemble des prix d'ouverture se réduit à un seul prix. La règle numéro 1 est discriminante et le marché ouvre au prix p* = a* = b*.

Lorsque a* > b*, la règle numéro 1 de maximisation du volume échangé n'est plus discriminante. Si le prix d'échange précédent est inférieur à b*, l'échange se fera à b*. Si il se trouve entre b* et a*, l'échange aura lieu au même prix que précédemment. Enfin, si ce prix est supérieur à a*, l'échange aura lieu à a*.

Prix multiples

Continu :

Si juste après l'arrivée d'un ordre de vente on se retrouve dans le cas de la figure (du bas), on voit que il y a des échanges possibles pour les prix p1 et p2. Une quantité Q1 va donc être échangée au prix p1 et Q2 au prix p2 (p1 n'est pas un prix d'équilibre).

Autre cas :

1) On apparie en continu mais on ne dévoile le résultat qu'à dates fixes (plus d'investisseurs participent à l'échange, plus de volume échangé, prix 503 et 504).
2) On peut récupérer tous les ordres qui se situent en dessous du point A et redérouler dessus un continu (même volume échangé, prix 502 à 504). 3) Oublier les priorités, et apparéier les ordres du plus contraignant au moins contraignant (plus de volume échangé, prix de 502 à 504).

Dates d'échange

- Appariement à dates fixes :
  les échanges ont lieu à des dates prédéfinies de la journée (toutes minutes TAIEX, 2 fois -11h30 et 16h30- ou une fois -15h- par jour : Paris).
- Appariement en continu :
  il y a appariement dès qu'il existe en carnet une contrepartie. Ex : l'arrivée d'un ordre d'achat à prix limite p déclenche une transaction si il existe en carnet un ordre de vente de prix limite inférieur ou égal à p, c'est-à-dire si :

  $$p \geq a_t.$$

  les échanges ont lieu dès que possible. Les dates d'échange sont endogènes et pas égales entre elles.
- Appariement volume en attente :
  il y a appariement dès que suffisamment d'ordres sont en carnet : dès que le volume cumulé des ordres en attente à l'achat et à la vente dépassent un certain volume v.

  $$\tau_{t+1} = \tau_t + \inf\left\{ \tau : D_{\tau_t}\left(\underline{p}\right) + \sum_{t=\tau_t+1}^{\tau_t+\tau} d_t\left(\underline{p}\right) \geq v, S_{\tau_t}\left(\overline{p}\right) + \sum_{t=\tau_t+1}^{\tau_t+\tau} s_t\left(\overline{p}\right) \geq v \right\}$$

  les échanges ont lieu dès que suffisamment d'ordres sont arrivés sur le marché. Les dates sont endogènes.
- Appariement volume échangeable
  il y a appariement dès qu'une quantité suffisante est échangeable

  $$\tau_{t+1} = \tau_t + \inf\left\{ \tau : Q^*_{\tau_t+\tau} \geq v \right\}$$

  les échanges ont lieu dès qu'un certain volume est échangeable. Là encore les dates sont endogènes et inégales.

Comparaison théorique : échange à prix unique

- Appariement séquentiel:

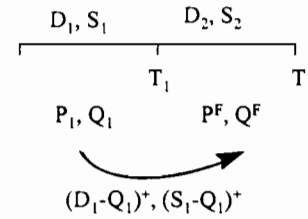
  entre 0 et T_1 ➡ P_1 et Q_1

  reste : $(D_1\text{-}Q_1)^+$ et $(S_1\text{-}Q_1)^+$

  $D^F = D_2 + (D_1\text{-}Q_1)^+$

  $S^F = S_2 + (S_1\text{-}Q_1)^+$

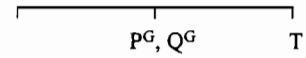
  ➡ P^F et Q^F

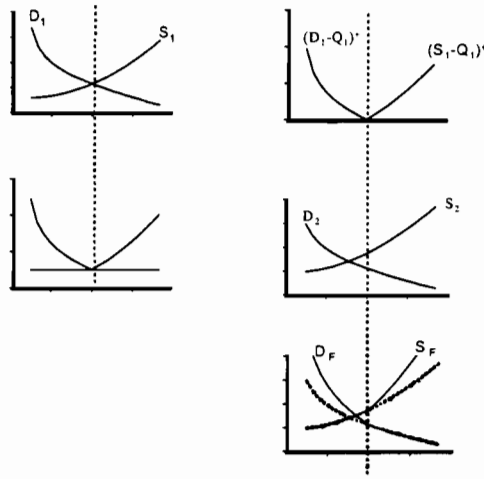


- Appariement global : on cumule les ordres arrivés entre 0 et T.

  ➡ P^G et Q^G

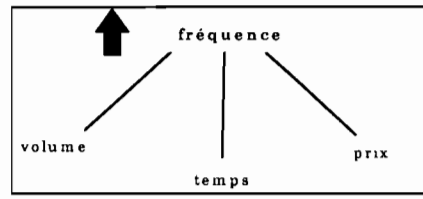
$D^G = D_1 + D_2$, $S^G = S_1 + S_2$

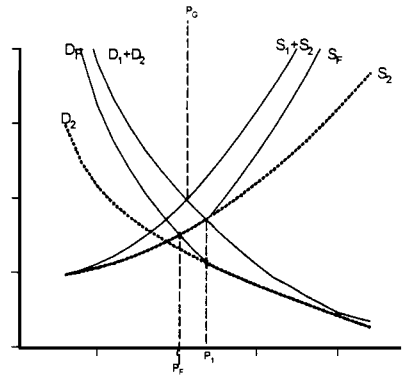




Propriété

- Le prix PG est dans l'intervalle [P1, PF]
- QG < Q1+QF
- Si P1 < PF, tous les vendeurs, servis en T par un appariement global, le sont aussi si le marché fonctionne en séquentiel.





Comparaisons simulées

1. Processus des arrivées

$$\begin{pmatrix}\log\beta_t\\ \log y_t\end{pmatrix} = \mathrm{N}\left(\begin{pmatrix}m_2\\ \mu_2\end{pmatrix},\Sigma_2\right)\begin{pmatrix}\log\alpha_t\\ \log x_t\end{pmatrix} = \mathrm{N}\left(\begin{pmatrix}m_1\\ \mu_1\end{pmatrix},\Sigma_1\right),$$

$$d_t(p) = x_t^*.1\mathrm{I}_{p\le\alpha_t^*}, \quad s_t(p) = y_t^*.1\mathrm{I}_{p\le\beta_t^*}$$

2. Effet de la fréquence de transaction

(I) appariement à dates fixes ; 1mn, 2mn, ..., 7mn,
(II) appariement volume en attente ; 100 000, 110 000, ..., 200 000 titres,
(III) appariement volume échangeable; 200, 400, ..., 1000 titres.

On échantillonne toutes les deux minutes.

Comparaisons : (i) volume global / jour

$$V = \sum_{n=1}^{N} v_n$$

(ii) volatilité des rendements

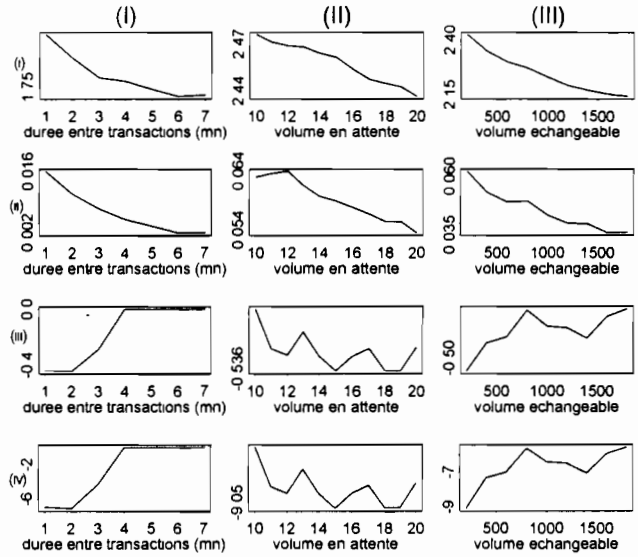
$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N}(r_n-\bar{r}_n)^2$$

(iii) autocorrélation d'ordre 1 des rdts
(iv) statistique de Student

Liaisons : fréquence - volatilité et volume

$$log\ y_n \approx \alpha + \beta\, log\ c_n$$

(I) appariement à dates fixes , (II) appariement volume en attente, (III) appariement volume échangeable



(i) volume échangé (iii) corrélation d'ordre 1

(ii) volatilité (iv) statistique de student

- Comme attendu, une augmentation de la fréquence d'échange augmente le volume échangé à la journée ainsi que la volatilité des rendements.
- Si l'on double la fréquence d'échange pour l'appariement à date fixe conduit à un accroissement relatif de 1.04 % du volume et de 2.41% de la volatilité des rendements (tableau 4.1 page 20).

$$\log y = \alpha + \beta \log c$$
$$\log y^* = \alpha + \beta(\log c + \log 2)$$
$$\log y^* = \log y + \beta \log 2$$
$$\log \lambda = \beta \log 2$$
$$\lambda = 2^\beta$$

- autocorrélation et statistique de Student de non corrélation des résidus sont différents selon la méthode. corrélations toujours non positives. Tests d'efficience de marché fondés sur le Student conduisent à rejeter l'hypothèse d'efficience de marché et les p-valeurs sont fortement dépendantes du mode de gestion et de la fréquence de transaction. Pas contraire à l'efficience de marché. En fait, les investisseurs adaptent leur comportement de façon a atteindre l'efficience et un comportement de placement d'ordres conduisant à l'efficience pour un mode perd cette propriété lorsque ce mode est modifié.

3. Effet du pas d'échantillonnage

(A) appariement en continu (à prix multiples)

(B) appariement volume en attente (v = 100 000 titres)

(C) appariement volume échangeable (v = 1000 titres)

(D) appariement toutes les 2 minutes

On échantillonne toutes les 1mn, 2mn, 5mn et 10mn

Comparaisons : autocorrélations d'ordre 1 corrigées,
autocorrélations d'ordre supérieur,
corrélations instantanées.

Si le rendement échantillonné à la minute r_t est autorégressif d'ordre 1 :

$$r_t - c = \rho(r_{t-1} - c) + \varepsilon_t$$

le rendement échantillonné toutes les h minutes

$$r_{ht}^{h} = r_{ht} + r_{ht+1} + \Lambda + r_{ht+h-1}$$

La corrélation d'ordre 1 vaut :

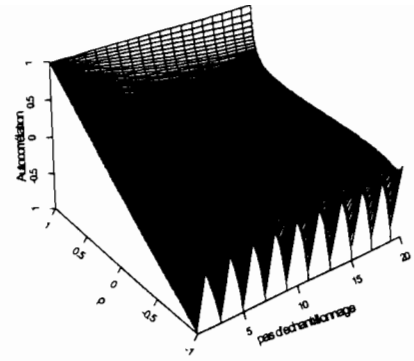
$$Cor\left(r_{ht}^{(h)}, r_{h(t+1)}^{(h)}\right) = \frac{\rho\left(\rho^h - 1\right)^2}{2\rho^{h+1} - h\rho^2 - 2\rho + h}$$

Méthode	autocorrélation			
	échantillonnage			
	1 minute	2 minutes	5 minutes	10 minutes
(A)	-0.43	-0.57	-0.37	-0.58
(B)	-0.45	-0.52	-0.52	-0.56
(C)	-0.51	-0.41	-0.59	-0.35
(D)	0	-0.45	-0.33	-0 31

Méthode	autocorrélation (à 10 minutes)			
	échantillonnage			
	1 minute	2 minutes	5 minutes	10 minutes
(A)	-0.05	-0.14	-0.12	-0.58
(B)	-0 05	-0.12	-0.12	-0.56
(C)	-0.06	-0.08	-0.12	-0.35
(D)	0	-0.09	-0.11	-0.31

Dans le premier tableau nous donnons les valeurs des corrélations empiriques pour les divers pas d'échantillonnage. On sait que le processus d'arrivée des ordres est Poissonien et donc Markovien. Est-ce que cette propriété se transmet aux rendements malgré le fait que les rendements sont issus des processus initiaux par des transformations non linéaires complexes dues aux modes d'appariement.

- Autocorrélation d'ordre 1 très sensible au pas d'échantillonnage et à la méthode de négociation.
- si rt échantillonné à la minute suit un AR(1), rt échantillonné toutes les h minutes suit un ARMA d'ordre supérieur.
- L'hypothèse rt 1mn suit un AR(1) est refusée quelque soit la méthode. Les dynamiques sont plus complexes pouvant présenter des dépendances d'ordre supérieur à 1 et des effets non linéaires dus à la non linéarité des procédures d'appariement.
- Fonction d'autocorrélation est le résumé des dépendances linéaires. Certains modes de gestion peuvent créer artificiellement des dépendance à des délais assez longs (+ d'une demi heure).



		corrélation instantanées			
		échantillonnage			
		1 mn	2 mn	5 mn	10 mn
(A)	rdt-volume	-0.27	**0.02**	0 04	0.01
	nb dates-rdt	-0.01	-0.01	0.16	0.19
	nb dates-volume	0.34	0.27	0.40	0.47
	fourchette-rdt²	**0.06**	0 11	0 08	0.39
	rdt-rdt²	-0.03	0.08	0.00	0.19
	fourchette-rdt	-0.07	-0.10	0.06	0 07
	nb dates-rdt²	0 05	0 09	0 01	0.08
(B)	rdt-volume	0 01	0 04	0 09	-0.01
	nb dates-rdt	0.00	-0.04	0.08	0.15
	nb dates-volume	0.29	0.20	0.21	**0 22**
	fourchette-rdt²	0.16	0.19	0 13	0 46
	rdt-rdt²	-0.05	0.08	0.10	0 16
	fourchette-rdt	-0.08	-0 08	0 01	-0.04
	nb dates-rdt²	0 00	-0 04	0.08	0 15
(C)	rdt-volume	0.01	-0.01	-0.02	-0 11
	nb dates-rdt	-0 03	-0.01	-0.02	-0.15
	nb dates-volume	0.60	0.56	0 64	0 74
	fourchette-rdt²	0.20	0.34	-0.07	0 11
	rdt-rdt²	-0.06	-0.13	0.11	0.18
	fourchette-rdt	-0.08	-0.12	-0.04	-0 08
	nb dates-rdt²	0 04	0 04	0 22	0.19
(D)	rdt-volume	0.03	0 06	-0 08	-0 08
	nb dates-rdt	0 00	NA	-0 13	NA
	nb dates-volume	0.79	0.27	0.49	0.17
	fourchette-rdt²	0.50	0.15	0 00	NA
	rdt-rdt²	-0.16	-0.19	-0.11	0.23
	fourchette-rdt	0.00	-0.14	0 00	NA
	nb dates-rdt²	0 50	NA	0.09	NA

On s'intéresse aux corrélations instantanées entre les diverses caractéristiques de marché : liaisons directes entre modification de prix, volume traité et fréquence d'échange, des liaisons entre plusieurs mesures de volatilité, des évaluations de prime de risque par l'effet sur les rdts des mesures de volatilité et enfin des liens liquidité et volatilité.

- rdt-volume :Les relations rendement-volume peuvent changer de signe : continu passe de -0.27 à 0.02. N'évolue pas dans le même sens : négative puis négligeable en continu, positive puis négative pour les autres.
- fourchette-rdt² : fourchette ne semble pas être une bonne approximation de la volatilité pour tous les appariements. de 0.5 en appariement toutes les 2 minutes (1mn) elle passe à 0.2 appariement volume échangeable, 0.16 appariement volume en attente et est négligeable en continu.
- rdt-volatilité : prime de risque peut être négative pour des petits pas

d'échantillonnage.

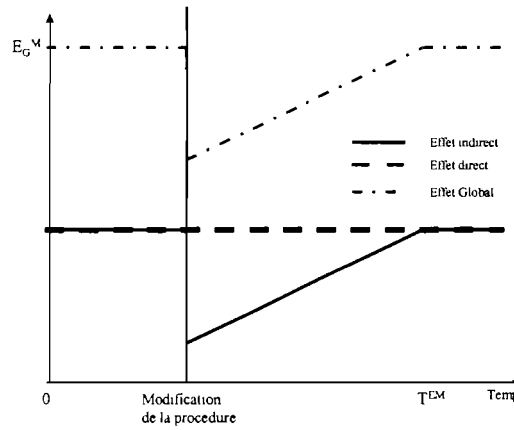
- volume-nb trans : les plus fortes (normal). vers 0.6-0.7 pour fixing et volume échangeable, 0.2-0.3 pour continu et volume en attente.

Conclusion

Une augmentation de la fréquence augmente le volume échangé et la volatilité des prix.

L'hypothèse d'efficience de marché est rejetée. Les investisseurs adaptent leur comportement pour atteindre l'efficience. Un comportement efficient pour un mode de gestion ne l'est plus pour un autre.

Les outils statistiques sont très dépendantes du pas d'échantillonnage et du système de négociation.



Données disponibles

Marché des actions :

Tous les titres échangés sur le marché français des actions depuis janvier 1990.

Données haute fréquence
données en temps réel de tout événement de marché . arrivée, annulation ou exécution d'un ordre

- Ordres	⇒	flux Topval, flux d'ordres BDM, meilleures limites, FMP
- Transactions	⇒	Table des transactions, échanges signés
- Evénements de cotation	⇒	Caractéristiques des valeurs, suspension, réservation, news (Reuters)
- Indices	⇒	Observations toutes les 30s

LES OUTILS DU MICRO MARKETING
Construction et exploitation d'un data warehouse pour le déploiement du marketing opérationnel

Jean-Michel Gautier
Professeur à HEC
Directeur général d'AXIS conseil

Groupe HEC - Département SIAD - 78351 JOUY EN JOSAS Cedex - France
Tél : 00 33 1 39 67 72 56 / Fax : 00 33 1 39 67 71 09
E-mail : gautier@hec.fr

Introduction

Le marketing des services, des produits de consommation et des biens d'équipement destinés aux particuliers, a connu depuis 50 ans différentes phases. Tout d'abord, le développement de la publicité et de la communication, à travers les mass média, accompagné d'une gestion des promotions et des prix nationale, puis l'apparition des premières segmentations de clientèle, la construction de gammes de produits et le ciblage des segments dans les plans média. Plus récemment, le développement et "l'explosion" des bases de données de clientèle et du "marketing de base de données" et enfin le développement depuis 5 ou 6 ans du marketing local à travers les outils du géomarketing.

Les parts de marché, que l'on conquérait il y a 15 ans à grand renfort de communication et de nouveaux produits ciblés, se gagnent à présent sur le champ de bataille du micro marketing (marketing local et marketing de base de données). A cela, plusieurs raisons :

- La qualité de la communication, des promotions et des plans média est aujourd'hui très bonne et l'accès à un niveau de qualité satisfaisant est relativement banalisé, ce qui rend la conquête d'un avantage concurrentiel, à travers ces outils, beaucoup plus onéreuse.
- L'innovation en matière de produits existe toujours, mais se voit fortement contrainte par des politiques industrielles, qui incitent plus à la concentration et à l'adoption de gammes limitées qu'à une créativité et à une exploitation d'un grand nombre de marchés de niches.

Conscientes de ces nouvelles données du marketing, les entreprises ont depuis une quinzaine d'années concentré leurs efforts sur une meilleure gestion de la relation aux clients, organisée à travers une communication personnalisée (marketing one to one) et une meilleure adaptation des produits dans les points de vente et les agences aux besoins des clientèles. Cette nouvelle forme de marketing cherche à adapter l'offre et la communication à chaque client, ou petit groupe de clients, et à créer ainsi une relation de fidélité aux produits, aux services et aux marques, cela s'appelle le micro marketing. Son principal moteur est l'idée que des parts de marché vont se gagner sur le terrain client par client, magasin par magasin, et non plus seulement de façon nationale, à travers un marketing produit.

Le micro marketing nécessitant de grosses bases de données d'information, des approches quantitatives lourdes et des méthodologies rigoureuses, le développement en est encore inégal selon les entreprises et les secteurs d'activité, souvent réservé aux

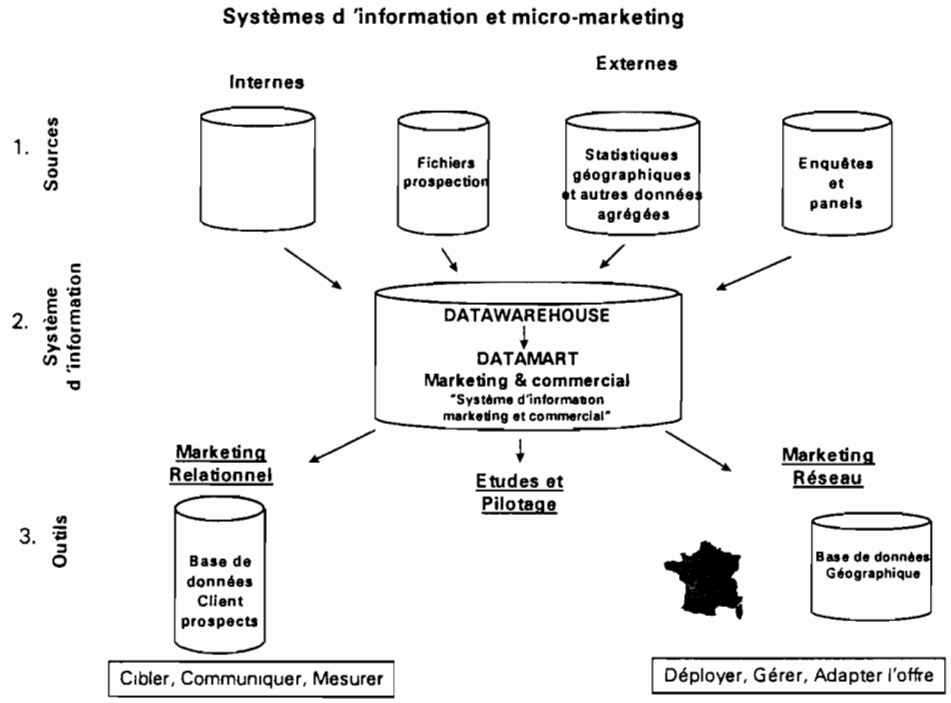
plus grosses entreprises. C'est précisément pour cette raison que les enjeux du micro marketing sont aujourd'hui très importants, car, en fonction de l'ampleur et des performances des solutions envisagées, les gains en parts de marché obtenus grâce à ces outils peuvent être considérables, et ce d'autant plus qu'un grand nombre d'entreprises aujourd'hui ne les utilisent pas ou en sont aux balbutiements dans leur développement.

Cette présentation a pour but de clarifier les enjeux et d'exposer les principaux outils nécessaires à la gestion d'un micro marketing efficace. Il existe, bien sûr, d'autres exploitations des systèmes d'information marketing, en particulier en terme d'études. Nous nous limiterons ici à la description des outils opérationnels les plus importants.

Cette présentation s'organise en 2 parties :

1. Les sources d'information et la construction du Datawarehouse
2. Les outils opérationnels du micro marketing

1- Les sources d'information et la construction du Datawarehouse



Le déploiement d'outils de micro marketing repose sur la capacité à connaître individuellement les clients et à décrire les territoires sur lesquels sont implantés les agences, les points de vente ou les forces de vente. La situation au regard de ces exigences est très différente d'un secteur d'activité à l'autre et d'un pays à l'autre.

La connaissance client est d'autant plus complète et exhaustive que les entreprises gèrent directement leurs ventes au client final. En effet, dans ce cas, les fichiers de gestion vont contenir toutes les informations client nécessaires à la gestion des ventes ainsi que toutes les informations sur les achats de ces clients au fil du temps. Lorsque ce n'est pas le cas, les entreprises tentent de se constituer des fichiers clients à travers différents procédés tels que :

- Cartes de crédit dédiées
- Cartes de fidélité, club
- Coupons de garantie
- Questionnaires sur les foires, salons, magasins
- Accords avec les distributeurs
- Achat d'adresses qualifiées dans des mégabases ...

Toutefois, en dehors des cartes de crédit ou de fidélité, ces différentes sources sont des photos instantanées et ne permettent pas de suivre les comportements d'achat du client dans le temps. Pour illustrer les différentes situations selon les secteurs d'activité, quelques exemples :

a) La Banque

Le banquier, pour la gestion des comptes, est amené à suivre le détail des opérations financières de ses clients (recettes, dépenses, comptes d'épargne, endettement) et recueille de façon réglementaire des informations socio-démographiques d'identification de son client (nom, adresse, date de naissance, profession, statut matrimonial...) voire plus lors de l'établissement de dossiers de prêts.

b) Les Télécommunications

L'opérateur de télécommunication a une connaissance détaillée de l'ensemble des relations de son client au travers des outils de télécommunication comme par exemple la localisation des numéros appelés et appelants, la durée des communications, l'utilisation des fax et modems. Il connaît souvent l'équipement téléphonique de son client, son adresse et quelques informations socio-démographiques requises à l'ouverture de la ligne.

c) L'Assurance

Selon les contrats détenus par le client, l'assureur va pouvoir identifier le parc automobile, la sinistralité ou bien le patrimoine immobilier et mobilier, ou les dépenses de santé auxquels s'ajoute l'ensemble des informations sur les différents individus assurés.

d) La Vente par correspondance

Le vépéciste connaît le détail des achats réalisés par ses clients, les moyens de paiement utilisés, le prénom, le nom et l'adresse.

e) La Presse

L'éditeur connaît les revues auxquelles s'abonne le client ainsi que son nom, son prénom et son adresse.

f) La Grande Distribution

Le distributeur, à travers la mise en place de cartes magasins ou de cartes de fidélité, identifie le détail des achats de ses clients ainsi que leur nom et leur adresse et souvent quelques informations socio-démographiques requises à l'ouverture de la carte.

g) L'Automobile

Le constructeur connaît les caractéristiques du véhicule acheté, la date, souvent le véhicule précédent, les services annexes souscrits, les réparations et entretiens effectués dans le réseau, le nom, le prénom et l'adresse du client.

Dans l'ensemble de ces exemples, la connaissance apportée par les fichiers de gestion est une connaissance instantanée dont le stockage est limité dans le temps aux besoins réglementaire des systèmes administratifs et comptables. L'enjeu principal dans la construction d'un datawarehouse orienté marketing est l'historisation de cette information et sa structuration.

Par exemple, dans le cas de la banque, la connaissance individuelle des opérations sur les comptes bancaires présente peu d'intérêt si ces dernières ne sont pas regroupées en agrégats significatifs (retrait par carte, paiement par chèque, retrait d'argent liquide, prélèvement de prêts, autres prélèvements...). De la même façon, dans la Vente Par Correspondance, ou la Grande Distribution, l'analyse des achats devra être structurée par marque, par famille de produits, par type de conditionnement, par niveau de gamme, par rayon... A ces données de systèmes comptables s'ajoutent des informations provenant des relations commerciales avec le client : mailings envoyés, visites de force de vente, réclamations clients.

Le dispositif sera en général complété par des informations extérieures issues de sources enquêtes, recensements, mégabase permettant de décrire plus en profondeur le comportement des clients et surtout d'apprécier ce comportement dans un univers de marché que l'entreprise ne peut connaître par ses seules données internes. Malheureusement, à quelques exceptions près, comme l'enrichissement à travers les mégabases, cette connaissance n'est acquise que sur un échantillon de clients ou de façon agrégée et devra donner lieu à des modélisations si l'on souhaite l'exploiter individuellement au niveau de chaque client ou prospect. Par ailleurs, les fichiers de prospection permettant le développement des activités sur de nouvelles clientèles sont rarement issues d'opérations internes mais viennent plutôt de fichiers achetés (annuaires, fichiers d'entreprises non concurrentes ou mégabase de consommateurs).

Le datawarehouse dans sa structuration et ses systèmes d'alimentation et de mise à jour doit permettre d'intégrer ces informations si possible au niveau de chaque individu ou foyer identifié par ses noms et adresses ou, lorsque cela n'est pas possible, sous forme de base de données agrégées décrivant des segments de clientèle ou des zones géographiques.

Le déploiement des réseaux, des points de vente ou agences, ou des forces de vente passent par la nécessité de disposer d'informations locales sur les clientèles résidentes ou migrantes des territoires où sont implantés ces points de vente. Le datawarehouse devra donc permettre d'une part de localiser et connaître les clientèles ou prospects existants que ce soit de façon individuelle ou à défaut sur un maillage le plus fin possible du territoire. Il devra également compléter cette connaissance client, prospect d'une description, là encore sur un maillage le plus fin possible, des foyers et individus résidants ou migrants sur la zone qu'ils soient ou non clients de l'entreprise (connaissance des marchés potentiels). Les sources permettant cette connaissance locale à un niveau de finesse suffisant sont soit les Instituts Statistiques nationaux à travers leur recensement, les fichiers administratifs et les grandes enquêtes, soit les mégabases de consommateurs.

Le datawarehouse devra gérer un système de maillage emboîté partant si possible du maillage le plus fin disponible dans le pays concerné et accumuler sur ces divers maillages les informations internes ou externes disponibles selon le niveau auquel elles peuvent être fournies. Il devra également permettre d'inscrire les clients et prospects à l'intérieur des zones, soit par des systèmes de géocodage utilisant des tables de voies pour identifier l'appartenance aux mailles, soit lorsque c'est possible par un géocodage de coordonnées (x, y) permettant l'affectation à n'importe quel maillage dès lors que l'on dispose des fichiers des contours de ces mailles. La validité statistique de l'information agrégée à l'intérieur des différentes zones de maillage est étroitement liée à la taille des

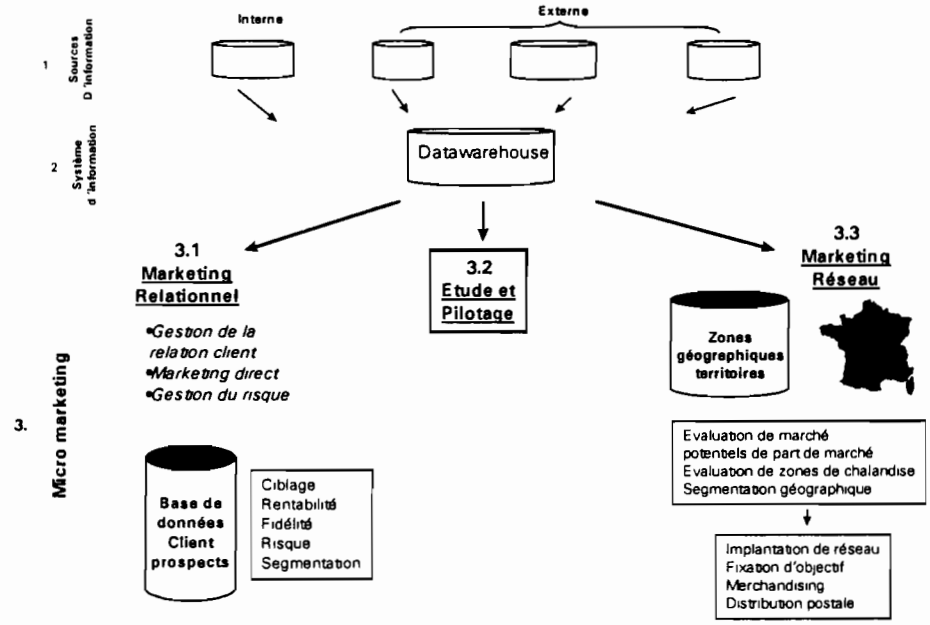
sources disponibles (nombre d'individus dans la zone) et la validité/précision des agrégats devra être prise en compte dans la construction du datawarehouse pour que les informations utilisées dans les modèles et les exploitations géomarketing aient une fiabilité suffisante. Ceci signifie que même si l'information est stockée de façon désagrégée ou sur un maillage très fin, des contraintes doivent être placées au niveau de son utilisation pour éviter des accès à des agrégats non significatifs. Tout une série de techniques de lissage et/ou de modélisation peuvent venir en renfort de l'estimation empirique pour autoriser l'utilisation d'agrégats sur des zones plus fines. Cette gestion de la structure géographique et de la fiabilité des estimations au sein du datawarehouse constitue le fondement d'un géomarketing efficace.

2-Les outils opérationnels du micro marketing

A partir du datawarehouse vont s'articuler différentes exploitations opérationnelles de l'information autour de la gestion de la relation individuelle avec les clients et prospects d'une part et autour du déploiement des réseaux d'agences/points de vente, ou des forces de vente, ainsi que de la fixation des objectifs et de l'adaptation locale des offres d'autre part. Les outils s'organisent selon le schéma suivant :

Il faut noter que le datawarehouse est aussi utilisé pour des études diverses en matière de marketing produit ou de marketing stratégique qui ne seront pas développés dans cette présentation.

La construction des outils du micro marketing



a) Le marketing relationnel

On peut structurer les besoins de connaissance de la clientèle et des prospects en 5 grandes catégories pour lesquelles devront être mis en place les outils de synthèse et de pilotage de l'action marketing.

Sensibilité produit et ciblage

L'observation des actions commerciales passées, la réalisation de tests, l'observation des clients déjà détenteurs ou acheteurs d'un produit, ou l'utilisation d'enquêtes spécifiques, permet de constituer pour un produit spécifique, ou une gamme de produits, des échantillons d'acheteurs réels ou potentiels au sein de la base de données clients/prospects. La comparaison de ces acheteurs aux non acheteurs permet d'établir un modèle de sensibilité d'un client/prospect à une offre pour un produit, ou une famille de produits, qu'il ne détient pas encore. Les clients/prospects dont la sensibilité sera la plus élevée constituent la cible privilégiée pour la commercialisation de ces produits. Les techniques qui permettent d'évaluer la sensibilité d'un client à une offre sont relativement nombreuses et conduisent toutes à une mesure de probabilité d'achat ou, à défaut, à un indice de potentiel qui permet de classer la population en fonction de son intérêt potentiel pour l'offre en question.

Parmi les différentes méthodes, les plus utilisées sont :

- *L'arbre de segmentation*
  Il répartit la population étudiée en segments le plus contrasté possible en matière de taux d'achat constaté sur la base des informations existant dans la base de données.
- *Le scoring*
  Système de notation des individus de la base utilisant leurs caractéristiques pour attribuer une note d'autant plus forte que l'individu est plus sensible à l'offre. Dans ce système, chaque information pertinente sur le client apporte sa contribution à la note globale. La cible à prospecter pour l'offre considérée est constituée des individus qui obtiennent les plus fortes notes dans le système de scoring.
- *La régression logistique*
  Il s'agit d'un modèle fonctionnant sur le même principe que le scoring mais sur la base d'une combinaison multiplicative de l'influence des variables et dont le résultat est une probabilité d'achat du produit concerné. La cible est constituée des individus qui présentent la plus forte probabilité d'achat.
- *D'autres méthodes*
  telles que les réseaux de neurones ou des techniques d'analyse discriminantes par estimation de densité fournissent également des solutions à ces problèmes de ciblage en classant les individus de la base clients/prospects selon des probabilités d'achat ou des indices de sensibilité par rapport à l'offre.

Toutefois, les 3 premières méthodes qui fournissent en général des résultats satisfaisants sont le plus souvent préférées en raison de la lisibilité, de l'interprétabilité de leurs résultats. Il faut également noter que dans la mise au point de ce type d'outils, on privilégiera toujours la prise en compte d'information client présentant (par expertise) un lien direct de causalité avec le phénomène étudié (c'est à dire l'achat de produit).

Fidélité de clientèle

Partant de l'analyse qu'il est en général beaucoup moins onéreux de conserver un client existant que d'en conquérir un nouveau, la lutte contre l'attrition de la clientèle est devenue un enjeu essentiel pour toutes les entreprises. La mise en place de programmes de fidélisation a lieu dans presque tous les secteurs d'activité et l'allocation des moyens ainsi que le choix des outils de fidélisation optimisés à partir d'une analyse prévisionnelle des risques individuels de départ et des causes potentielles de cette infidélité. La mesure de ce risque et l'identification des causes et un élément fondamental de la mise en place des programmes de fidélisation. Dans tous les secteurs

où l'activité est récurrente, soit de façon contractuelle à travers des liens de prestations de services (comme la banque ou l'assurance ou des abonnements comme pour la téléphonie, la presse ou encore par le jeu des achats répétés dans tout le secteur de la grande consommation), la fidélisation est au cœur des préoccupations. Les outils dont nous allons parler ci-après concourent à la mise en place de ces programmes en conjonction avec des analyses de la rentabilité client dont nous parlerons plus loin, et avec le test et l'évaluation des coûts des actions de fidélisation envisagées.

La détermination du risque d'attrition individuelle s'effectue sur la base client à l'aide des informations disponibles dans cette dernière. A partir de l'observation des clients infidèles sur la dernière période, on cherche à travers un modèle, comme pour l'estimation de la sensibilité produit, à évaluer la probabilité de départ d'un client donné à un instant donné. Toutefois, l'analyse de la fidélité est plus complexe que le ciblage même si les modèles statistiques utilisés sont de même nature, dans la mesure où il convient :

- d'une part, de distinguer l'analyse structurelle de fidélité qui donne la probabilité pour un certain type de clients de partir sur un horizon de 6 mois ou 1 an, et l'analyse conjoncturelle fondée le plus souvent sur la détection d'événements tels que réclamations, baisse de la consommation, etc., qui fournit plutôt une alerte sur un risque de départ à très court terme
- d'autre part, parce que en matière de fidélité, les clients doivent être souvent analysés par cohorte en fonction de leur ancienneté, dans la mesure où les recrutements de clientèle d'une entreprise sont en général variables dans le temps en fonction de la maturité de la marque et des produits.

A cette mesure du risque, il est souvent utile d'adjoindre, lorsque l'on a détecté un client à risque, une analyse des facteurs déclenchant potentiels les plus importants, car ils peuvent constituer des indications pour des leviers d'action dans les programmes de fidélisation. Les techniques statistiques utilisées dans ce cadre vont de la segmentation à de simples tests de comparaison de proportions ou de moyennes.

Le risque financier

Cet aspect de l'analyse des comportements de la clientèle n'est pas présent dans toutes les activités. Néanmoins, il peut être essentiel dans certains secteurs comme la banque en matière d'attribution de prêts, ou comme la Vente Par Correspondance, ou la télévision par abonnement, ou encore la téléphonie mobile, etc., quand les biens ou les services sont facturés après avoir été livrés. L'analyse du risque se traite à partir de l'observation des incidents sur une période récente en comparant des bons clients à des mauvais clients et en établissant des modèles d'évaluation de la probabilité d'incident avec les mêmes techniques statistiques que pour l'évaluation de la sensibilité produit. Certains types de risque sont néanmoins plus complexes, lorsqu'il s'agit non pas d'une prise de décision uniquement tel que l'ouverture d'un dossier ou l'attribution d'un prêt, mais qu'il est question, par exemple, dans une banque de suivre au quotidien les incidents sur les comptes. Des méthodes statistiques particulières doivent être alors utilisées que nous ne détaillerons pas ici.

b) Segmentation de clientèle

Pour une meilleure efficacité de la démarche marketing, il est important de distinguer parmi ces clients différents segments regroupant des individus ou des foyers dont les comportements d'achat et les caractéristiques se ressemblent, et ce afin d'adapter pour chacun de ces segments l'offre de produit et la communication autour de cette offre. Les démarches de segmentation sont aujourd'hui généralisées dans la plupart des secteurs, soit en segmentant directement des clients dans la base de

données, soit en construisant des segments à partir d'enquêtes. Selon l'usage que l'on souhaite faire de cette segmentation, l'accent sera plutôt mis dans la construction des segments sur l'homogénéité en terme socio-démographique, quitte à recouvrir une certaine diversité de comportements dans les segments, ou sur l'homogénéité comportementale, quitte à avoir une certaine diversité socio-démographique dans les segments. Les méthodes statistiques utilisées dans la construction des segments sont les typologies ainsi que la segmentation multicritères et le plus souvent un mixte des deux.

c) Mesure de la rentabilité client

Une bonne allocation des budgets consacrés aux différentes actions marketing passe par la mesure des enjeux individuels client par client, c'est à dire, une évaluation de la rentabilité actuelle et future d'un client. Ces techniques, plus couramment connues sous le nom de "lifetime value" consistent à combiner une évaluation des rentabilités actuelles utilisant des éléments de comptabilité analytique avec des prévisions d'évolution de la situation du client (y compris départ du client) pour estimer sur une période de temps de plusieurs années la rentabilité que l'on peut attendre sur les achats des produits et des services de ce client. La mise en place de ces indicateurs de rentabilité sur la base de données client passe d'une part par une évaluation comptable des résultats par activité ou par produit, d'autre part, par la modélisation probabiliste des cycles de vie client et des taux d'attrition.

d) Le marketing local ou géomarketing

Pour beaucoup d'entreprises qui utilisent peu ou pas le marketing direct et les outils dont il a été question précédemment, l'essentiel des ventes se fait :

- à travers un réseau de distribution (magasins, points de vente, agences) qui peut être soit un réseau propriétaire (appartenant à l'entreprise) soit un réseau de distribution multimarques
- ou encore à travers une force de vente chargée de démarcher la clientèle

Dans tous les cas, la distribution est donc organisée autour de territoires de vente qui sont : soit les zones de chalandise des points de vente, soit les territoires attribués aux forces de vente.

Pour optimiser les ventes réalisées à travers ce réseau de distribution, différents leviers d'action sont possibles :

- l'optimisation de l'implantation des points de ventes ou de la force de vente
- l'adaptation de l'offre à la clientèle du territoire ou de la zone de chalandise
- la fixation d'objectifs commerciaux pour les points de vente ou la force de vente

La mobilisation de ces leviers d'action nécessite un système d'informations permettant de localiser les clientèles, de connaître les zones de chalandise des points de vente ou les territoires des vendeurs et de décrire les caractéristiques des populations résidentes ou migrantes sur ces zones, tant en terme socio-démographique, qu'en terme de comportement d'achat.

Les outils mis à contribution à cet effet sont :

- représentation cartographique
- localisation, et calculs de temps de parcours
- informations sur les zones
- méthodes d'estimation de potentiels

Représentation cartographique

Elle se fait, en général, à l'aide de logiciels informatiques SIG (Système d'Information Géographique) qui ont pour fonctionnalité la représentation sur des cartes de l'information ponctuelle ou d'informations par zone. Ces SIG sont, en général, conçus pour accepter différentes sortes de cartes :

- des cartes dites vectorielles lorsque tracés des rues, des voies, des fleuves, et les contours des zones sont stockés sous forme de coordonnées de segment,
- des cartes dites raster quand elles sont stockées sous forme d'ımages photographiques de cartes papier.

Les cartes vecteurs ont, en général, l'avantage de permettre des changements d'échelle faciles et d'occuper beaucoup moins de place dans les fichiers informatiques, par contre les cartes raster sont plus agréable à lire.

Le niveau de précision des cartes est un paramètre essentiel de l'analyse géographique dans la mesure où il détermine avec quelle finesse on peut représenter l'information disponible. Les cartes les plus précises vont jusqu'à donner le dessin des rues avec l'identification des carrefours et les numéros des immeubles à chaque carrefour.

La représentation géographique n'est pas indispensable pour le géomarketing, néanmoins le support visuel constitue une aide essentielle à l'interprétation de l'information géographique.

Les logiciels de SIG les plus diffusés sont au nombre de 2, ARCVIEW (de la société ESRI) et MAPINFO. Ces logiciels permettent, sur la base de fonds de cartes achetés auprès de sociétés spécialisées, de superposer des symboles représentant des clients ou des magasins localisés à leur implantation exacte, ou de colorier, de placer des histogrammes ou différents symboles sur un découpage du territoire (cartographie dite thématique). Des fonctions annexes permettent, lorsqu'on fait l'acquisition des données nécessaires, de représenter la zone dite isochrone qui se trouve à un certain temps de parcours d'un point donné (point de vente par exemple). On s'en sert , par exemple, pour fabriquer des zones de chalandise théoriques de point de vente.

Localisation

L'analyse des caractéristiques d'un territoire à partir d'un découpage donné en zones (maillage ou zonage) suppose que l'on puisse localiser et affecter à ces zones les clients de la base client et le cas échéant décrire ces zones par des informations extérieures. Cette opération de localisation s'appelle le géocodage. Elle peut s'effectuer de 2 façons :

- *Le géocodage en coordonnées (x, y) des adresses des clients*
  Ce géocodage nécessite pour l'ensemble du pays une cartographie à la voie permettant de positionner exactement une adresse sur une carte (y compris le numéro de l'immeuble). Selon les pays, les fonds de cartes nécessaires existent ou n'existent pas et sont plus ou moins onéreux. Le géocodage (x, y) présente l'énorme avantage de permettre l'affectation des points à un zonage quelconque dès lors qu'on dispose des contours des zones. Les principaux SIG proposent cette fonctionnalité. Il existe par ailleurs des logiciels spécialisés, qui ne nécessitent pas l'acquisition d'un SIG, mais qui travaillent directement sur les bases de données pour affecter les individus aux zones.
- *Le géocodage direct à la zone*
  Ce dernier se fait à partir de tables informatiques décrivant les communes et les voies ou les portions de voies incluses dans les zones. La recherche de l'adresse du client dans ces tables permet de l'affecter à l'une des zones du maillage. Le principal inconvénient de cette technique est qu'il faut choisir à l'avance un zonage pour fabriquer l'outil de géocodage.

Dans la pratique, l'analyse de clientèle peut être réalisée sur des maillages spécifiques à l'entreprise (zone commerciale, territoire, zone de chalandise). Mais, dès lors que l'on souhaite raccrocher des données externes de description de marché aux données de clientèle, on peut être contraint à l'adoption de certains maillages imposés par le fournisseur de données. Ainsi par exemple en France, lorsque l'on souhaite obtenir les données du recensement, on doit adopter pour certaines informations le découpage îlot, pour l'autre le découpage IRIS.

Informations sur les zones

Comme on vient de le voir, le choix d'un zonage de travail est un élément clef pour la constitution d'une base d'informations géographiques. Les principales contraintes sur le choix du maillage sont la disponibilité d'informations externes d'une part, et d'autre part la fiabilité de l'agrégation statistique des informations à l'intérieur de chaque zone. Dans la pratique et sauf considérations budgétaires, on cherchera à se constituer des maillages hiérarchiques en rattachant à chaque niveau d'agrégation les informations disponibles. Puis, lors de l'analyse, on choisira le niveau qui convient le mieux dans un compromis fiabilité/finesse de l'analyse, quitte à garder constantes des informations qui ne sont fiables que sur un niveau supérieur.

Méthodes d'estimation de potentiels

La plupart du temps, les informations sur les marchés (parts de marché, marché total, potentiel) ne sont pas disponibles sur des sources présentant un bon niveau de fiabilité à l'échelon local, à l'exception de quelques informations de parts de marché qui peuvent être obtenues à travers les mégabases. Par contre, ces informations sont en général disponibles au niveau national dans des enquêtes ou panels. Pour pallier cette difficulté, on procède souvent, à partir des dites enquêtes, à la construction de modèles statistiques qui permettent de reconstituer correctement les valeurs locales de ces parts de marché ou potentiels.

Ces modèles sont en général d'autant plus fiables et précis que les phénomènes étudiés ne présentent pas trop de particularismes locaux autres que dus aux structures de clientèle. De ce point de vue, en particulier, les modélisations de la demande (estimation de marchés potentiels) sont en général plus fiables que les modélisations potentiellement liées à l'offre (parts de marché). Bien entendu, la qualité des résultats dépend des méthodes de modélisation utilisées, de la qualité et de la quantité des informations disponibles pour construire ces modèles, et de la taille des panels utilisés.

A l'aide de tous ces outils, et au travers d'analyses statistiques et de règles d'expertise, on va chercher à répondre aux différents objectifs du géomarketing, c'est à dire :

- à travers la comparaison de la pression commerciale avec les potentiels de marché par zone, restructurer les réseaux pour adapter leur déploiement aux situations des marchés et intégrer dans la réflexion la prise en compte de l'implantation de la concurrence. Puis fixer à chaque entité du réseau commercial les objectifs détaillés par produit et segment de clientèle en fonction des marchés potentiels et de la pression concurrentielle
- à travers la comparaison de l'offre actuelle des magasins avec la structure de demande de la zone de chalandise, adapter l'offre, le linéaire et plus généralement le dispositif du merchandising aux besoins des clientèles.
- gérer la communication et en particulier le marketing direct pour générer du trafic vers les points de vente en ciblant les adresses à plus fort potentiel à l'intérieur des zones de chalandise.

LA MESURE D'AUDIENCE DES MEDIAS

Anne-Marie DUSSAIX,
Professeur à l'ESSEC

Ecole Supérieure des Sciences Economiques et Commerciales
B.P.105 - 95021 Cergy-Pontoise Cedex - France
tél : (33) 1 34 43 30 74 - fax : (33) 1 34 43 30 01
mail : p_dussaix@edu.essec.fr

Dans le monde entier, les enjeux liés aux médias sont considérables. Les enjeux liés à la mesure d'audience des médias le sont également.

En France, par exemple, en 1997, le montant total des investissements plurimédia des annonceurs dans la presse, la télévision, l'affichage, la radio et le cinéma avoisinait les 70 millions de francs.

Les enjeux de la mesure d'audience sont importants pour les médias eux-mêmes. Les études d'audience permettent de quantifier et de qualifier le lectorat des différents journaux et magazines, les auditeurs des stations de radio et chaînes de télévision. Les mesures d'audience conditionnent les ventes d'espace publicitaire. Les études d'audience sont également importantes pour les annonceurs et leurs agences de publicité auxquels elle permet une allocation optimale de leurs ressources. Ces deux groupes d'acteurs ont des intérêts parfois divergents.

Dans cette présentation, nous décrirons les principales méthodologies d'enquête utilisées dans le monde pour mesurer l'audience des principaux médias, en illustrant certains aspects par des exemples tirés des études réalisées en France[1].

Les méthodes d'enquête dans ce domaine se caractérisent généralement par une recherche de qualité, qui s'explique en partie par les enjeux économiques importants de la mesure d'audience que nous avons déjà évoqués.

Cette recherche de qualité peut être illustrée par quelques observations générales :

- on constate tout d'abord une grande diversité des méthodes d'enquête utilisées, adaptées à chaque média et à ses objectifs : panel de foyers équipés d'audimètres pour l'audience de la télévision, enquêtes sur l'écoute de la veille ou carnets d'écoute d'une semaine pour la radio, ...;
- ces enquêtes se déroulent le plus souvent de façon continue et donnent lieu à un nombre généralement important d'interviews, permettant la mesure de l'audience sur tout l'univers ou sur des cibles plus fines ;
- elles font l'objet d'améliorations méthodologiques régulières qui sont le plus souvent testées avant implantation à l'intérieur des enquêtes existantes ou par des tests spécifiques ;
- enfin, les méthodologies d'études s'appuient en partie sur des "recommandations" exprimées au niveau mondial, dont l'objectif est l'harmonisation des différents

[1] La présentation des études d'audience en France n'est pas exhaustive. Le lecteur intéressé trouvera une description des études de référence en France dans le Guide des mesures d'audience CESP et dans le Médiagraphe 98/99 (CESP, 1999) et dans A.M. DUSSAIX (1999, à paraître) que cette présentation reprend partiellement.

systèmes de mesure d'audience de la radio et de la télévision (cf. European Broadcasting Union[2], 1997). Les différents documents publiés donnent aussi une description détaillée des différents systèmes de mesure utilisés en Europe pour la radio, et dans le monde pour la télévision.

L'organisation des enquêtes de mesure d'audience diffère selon les pays. En France, les études d'audience sont réalisées par des instituts de sondage et sont contrôlées par un organisme, le Centre d'Etude des Supports de Publicité (CESP) cofinancé à la fois par les médias, les annonceurs, les agences de publicité et les centrales d'achat. Le contrôle s'exerce sur la méthodologie de l'enquête, sur la collecte des données et sur les résultats. En France, dans le domaine des sondages, le seul autre organisme exerçant un contrôle sur la qualité des enquêtes est la Commission des Sondages pour les sondages préélectoraux, mais son contrôle s'exerce seulement *a posteriori*.

Avant de décrire les principales méthodes d'enquêtes utilisées pour chacun des médias, nous préciserons brièvement les utilisations principales des études d'audience.

1- Les études d'audience : pour quoi faire ?

Les études d'audience ont deux utilisations principales :

- la quantification et la qualification de ceux qui "fréquentent" le média : combien sont-ils ? et qui sont-ils ? Encore faut-il définir ce que l'on entend par "fréquentation", par lecteurs ou par auditeurs ;
- le média-planning, c'est-à-dire la sélection optimale des médias et des supports par les annonceurs et les agences de publicité.

Le premier objectif concerne directement les médias ; par les études d'audience, les médias peuvent quantifier et qualifier leur lectorat dans le cas de la presse, leurs auditeurs dans le cas de la radio et de la télévision, les utilisateurs de leur site dans le cas d'Internet, ... Les résultats de ces études permettent aussi aux médias de suivre le succès de leur contenu ou de leurs programmes et de déterminer les tarifs de publicité.

La deuxième utilisation des études d'audience est le média-planning, c'est-à-dire le choix optimal des médias et des supports adaptés à la cible visée.

2- Les études d'audience de la télévision

Depuis le milieu des années 80, la mesure de l'audience de la télévision par des audimètres s'est largement développée dans le monde et est encore actuellement le système le plus utilisé.

De nombreux pays utilisent donc l'audimètrie individuelle. Le principe de l'audimètrie individuelle consiste à équiper un panel (c'est-à-dire un échantillon permanent de foyers) d'audimètres à bouton poussoir. L'audimètre est un appareil qui enregistre "en continu" (en France, c'est à la seconde près) les différents états du récepteur (allumé – éteint) et qui détecte la réception des chaînes ou de tout appareil périphérique connecté. L'audimètre possède une télécommande qui permet aux individus (membres du foyer au-dessus d'un âge minimal et invités) de déclarer leur présence devant le récepteur et leur départ. L'appareil enregistre donc ainsi simultanément l'audience foyer et l'audience individuelle. Les données sont rapatriées la nuit vers l'organisme en charge de la mesure.

---

[2] European Broadcasting Union (Union Européenne de Radio-Télévision) – Ancienne Route 17 A , Case postale 67, CH-1218 Grand-Saconnex (Geneva) – Tél. (+41) (22) 71 72 111 – Fax (+41) (22) 71 72 481 – Email : ebu@ebu.ch

L'historique de la mesure en France donne un bon aperçu des étapes successives qu'a suivies cette mesure d'audience. De 1961 à 1981, l'enquête était réalisée en face à face au domicile de l'interviewé, avec un effectif total annuel de 12 000 à 14 000 interviews. Le questionnaire portait essentiellement sur les habitudes d'audience et sur les écoutes de la journée de la veille par tranche de quart d'heure. A partir de 1988, ce type d'étude a été remplacé par un panel audimétrique de foyers qui ne permettaient pas de rendre compte du nombre et de l'identité des individus auditeurs. C'est en Mars 1988 que l'institut Médiamétrie a mis en place un panel de foyers équipés d'audimètres à boutons poussoirs permettant ainsi simultanément la mesure de l'audience foyer et de l'audience individuelle.

Deux définitions différentes de l'audience individuelle sont utilisées dans les systèmes audimétriques :

- soit on demande aux membres du panel de déclarer leur présence dans la pièce où le téléviseur est allumé,
- soit on demande aux membres du panel de se déclarer lorsqu'ils se considèrent comme téléspectateurs.

Certains pays d'Europe utilisent une définition différente : ils demandent aux panélistes de se déclarer quand ils sont dans la pièce et en mesure de regarder la télévision.

Cependant, certaines expérimentations suggèrent que la mesure diffère peu selon la consigne donnée aux panélistes. L'avantage de la première définition est son caractère objectif. C'est cette définition qui est adoptée dans le panel Médiamat en France. Elle facilite ainsi les études coïncidentales internes qui consistent à interroger par téléphone un sous-échantillon de panélistes en demandant qui, dans le foyer, est dans la pièce du téléviseur au moment de l'appel. On vérifie ensuite, à titre de contrôle, si leur présence était bien déclarée par leur bouton – poussoir enclenché.

Le taux de rotation des foyers du panel est généralement compris entre 15 à 25 %, hors rotation forcée, c'est-à-dire remplacement des panélistes ayant atteint la durée maximale de participation fixée.

La qualité des résultats est très liée à de nombreux facteurs, parmi lesquels on peut citer en particulier :

- l'existence ou l'obtention d'informations de qualité sur les structures de la population équipée télévision, en particulier sur des variables démographiques et d'équipement audiovisuel corrélées au comportement d'écoute ;
- la vérification régulière et systématique de la structure du panel ;
- le suivi de l'assiduité des panélistes ;
- l'existence des procédures formalisées pour étudier les audiences nulles, ou détecter les audiences individuelles de durée longue.

L'avantage majeur de l'audimétrie est de réduire au minimum l'appel à la mémoire des panélistes. Cependant, ces audimètres sont relativement coûteux. Lorsque le coût des appareils et le coût de gestion du panel sont prohibitifs ou lorsque le transfert des données entre les foyers panélisés et l'organisme en charge de l'étude est difficile, l'UER recommande deux solutions alternatives :

- les panels avec carnet d'écoute et relevé des audiences des différentes chaînes de télévision au quart d'heure. Ces panels permettent d'obtenir des données longitudinales ;
- les enquêtes sur l'audience de la veille avec relevé par quart d'heure, soit en face à face, soit par téléphone ; cette deuxième méthode permet un plus grand échantillon que la méthode précédente mais ne permet pas d'étudier les

comportements dans le temps.

Actuellement, face aux changements considérables du paysage et des comportements télévisuels, des expérimentations sont en cours dans le monde pour tester de nouvelles technologies de mesure de l'audience de la télévision et en particulier de la télévision numérique.

3- La mesure d'audience de la radio

Théoriquement, l'audience de la radio pourrait être mesurée comme la télévision par l'intermédiaire de "radiomètres". Mais alors que la télévision est essentiellement regardée sur le récepteur à domicile, la radio est écoutée sur de nombreux postes et dans des endroits variés : domicile, voiture, lieu de travail. Installer des radiomètres sur tous ces postes est irréalisable et serait trop coûteux ; d'autant plus que la mesure d'audience des stations régionales ou locales nécessite des échantillons de taille très importante.

Toutefois, des radiomètres sous forme de montres bracelets, qui seraient portés par des panélistes, sont en cours d'expérimentations et pourraient être mis en service en France en 2001(cf. Gane,1997).

Actuellement, la mesure d'audience de la radio est encore faite par enquêtes. L'European Broadcasting Union a publié en 1997 la première édition de ses recommandations dans lesquelles il développe les deux techniques actuellement employées dans le monde :

- l'audience de la veille (24-hour Recall), soit par enquête en face à face, soit par téléphone ;
- le carnet d'écoute auto-administré sur 7 jours (7-day self completion diary) avec dépôt et collecte soit à domicile, soit par voie postale.

Les deux méthodes ont des objectifs sensiblement différents. Toutes deux sous-entendent une écoute consciente de la radio. Ce sont des méthodes complémentaires dans leur utilisation avec leurs avantages et inconvénients respectifs. Dans les pays européens, la première méthode est plus fréquemment utilisée. En France, les deux méthodes coexistent. La présentation du système de mesure d'audience français illustrera ces objectifs.

L'institut MEDIAMETRIE propose en effet deux outils complémentaires :

- L'enquête 75 000⁺ dont l'objectif est de mesurer les niveaux d'audience de la radio (et aussi de la télévision et la fréquentation du cinéma) grâce à un recueil par téléphone de l'audience veille auprès d'un échantillon renouvelé quotidiennement. Dans cette étude, le concept d'audience utilisé est *"l'écoute ne serait ce qu'un instant"* dans le demi-quart d'heure (au lieu de *"l'écoute ne serait ce qu'un instant"* dans le quart d'heure jusqu'à fin 1998). L'objectif est de décrire les comportements d'audience de la "radio en général" et des stations par des paramètres caractérisant des comportements quotidiens moyens (jour nommé, audience moyenne lundi au vendredi, week-end). L'enquête relève aussi l'écoute télévision de la veille relevée au quart d'heure près avec identification de la chaîne.
  Les résultats d'audience sont publiés 4 fois par an. En 1988, l'échantillon total annuel comprenait 75 000 individus de 15 ans et plus, à raison de 250 interviews par jour.
- La deuxième étude est faite par panel. Le panel radio a suivi en 1997/98, avec un carnet d'écoute auto-administré, 7200 personnes recrutées par téléphone auprès du réservoir que constitue l'enquête précédente 75 000. L'objectif est très

différent. Il s'agit d'observer l'évolution des comportements d'écoute radio des individus dans le temps, en particulier de mesurer les accumulations d'audience, c'est à dire la proportion d'individus qui ont écouté telle station, tel quart d'heure au bout de t jours. L'étude permet aussi de mesurer la fidélité aux stations et aux émissions.

Les recommandations déjà citées soulignent que la mesure d'audience de la radio est très sensible à la technique de mesure à cause des nombreuses situations dans lesquelles elle est écoutée avec des niveaux d'attention très variables.

Avec les radiomètres actuellement en cours d'expérimentation, la définition de l'écoute de la radio va évoluer puisque les radiomètres capteront tous les sons audibles par leurs porteurs, quel que soit le degré d'attention qui leur est porté.

4- La mesure d'audience de la presse

Un quotidien ou un magazine pouvant être lu par plusieurs personnes différentes, les études d'audience de la presse ont pour objectif d'estimer le nombre de lecteurs d'un titre donné au cours d'une période de temps donnée. Encore faut-il définir ce que l'on entend par lecteur. Remarquons que la presse est le seul média pour lequel existe une certaine validation de la mesure d'audience par la diffusion, c'est à dire le nombre d'exemplaires de journaux et de magazines vendus ou distribués gratuitement.

La plupart des études d'audience sur la presse dans le monde sont réalisées en face à face, méthode qui permet de présenter aux interviewés les logos des titres, généralement très nombreux.

L'audience des magazines et des quotidiens est mesurée assez généralement par l'indicateur "Lecture Dernière Période" (ou LDP) reconstituée à partir d'une question sur la date de dernière lecture. Sont dits lecteurs LDP ceux qui déclarent avoir "personnellement lu, parcouru, ou consulté un numéro, même s'il s'agit ancien, que ce soit chez vous ou ailleurs" la veille pour les quotidiens, il y a moins de 8 jours pour les hebdomadaires, il y a moins de 30 jours pour les mensuels, etc.

La LDP est de loin la mesure d'audience la plus utilisée dans le monde (cf. Readership measurement in Europe, 1996). Elle pose cependant un certain nombre de problèmes : elle fait largement appel à la mémoire et favorise ainsi les biais de notoriété des titres; elle confond les lectures répétées d'un même numéro avec la lecture sur une même période de plusieurs numéros consécutifs (un lecteur de plusieurs numéros d'un même hebdomadaire sur une seule semaine est compté plusieurs fois); par contre, un lecteur de plusieurs numéros le même jour est compté une seule fois.

D'autres définitions du lectorat sont aussi utilisées dans l'enquête et donnent lieu à mesure :

- les lecteurs 12 derniers mois sont ceux qui déclarent avoir "lu, parcouru, ou consulté le titre au cours des douze derniers mois" ; cette question sert de question filtre en début d'enquête ; les résultats sont très sensibles à la façon dont la question est posée car cette question fait trop appel à la mémoire des interviewés ;
- les lecteurs réguliers sont ceux qui déclarent avoir l'habitude de lire, parcourir ou consulter un hebdomadaire toutes les semaines, un bimensuel tous les 15 jours, etc.

La mesure d'audience de la Presse Quotidienne utilise aussi une autre définition de l'audience : la Lecture d'un Numéro Moyen (ou LNM), c'est-à-dire le nombre moyen de lecteurs par numéro calculé sur la base des 6 derniers numéros parus.

Afin d'illustrer l'impact évident de la définition utilisée sur la mesure, le cumul

juillet 1997 à juin 1998 de l'étude Presse Magazine[3] en France donne les résultats suivants pour trois hebdomadaires :

	Lecteurs Dernière Période	Audience 12 mois	Lecteurs réguliers
Paris-Match	9,6 %	56,4 %	4,6 %
L'EXPRESS	5,3 %	35,2 %	3,0 %
Femme Actuelle	18,4 %	58,7 %	12,4 %

Les tailles d'échantillon très importantes de ces enquêtes (cf. Le médiagraphe 98/99 où sont décrites les principales études en Europe) permettent aux éditeurs de qualifier leurs lecteurs à partir d'un nombre de questionnaires suffisant.

5- L'audience de l'affichage

L'affichage est le seul média ayant comme objectif unique de véhiculer des messages publicitaires. En affichage, la définition du contact est la plus large qui soit : c'est le passage devant le panneau, en tenant compte de la distance et de l'axe de vision.

Les études d'audience de l'affichage ont donc comme objectif de mesurer dans les différentes agglomérations les passages effectués devant les réseaux de panneaux d'affichage et de répondre aux questions suivantes :

- combien y a-t-il eu de passages devant un nombre d'emplacements précis pendant la durée d'une campagne d'affichage donnée dans une agglomération déterminée ?
- quel pourcentage de la population de cette agglomération a été touché par cette campagne ?
- combien de fois en moyenne ces passants ont-ils été touchés ? Quelle est la distribution des contacts ?

A la différence des autres médias, la mesure d'audience des campagnes d'affichage doit être faite ville par ville.

En France, les enquêteurs sont équipés de terminaux portables; ils recueillent essentiellement les habitudes de déplacement et l'ensemble des déplacements de la veille à l'aide d'un logiciel cartographique qui facilite la reconstitution de l'itinéraire emprunté et des moyens de transport utilisés, et ce, pour chaque déplacement déclaré réalisé la veille.

En conclusion, de nombreuses recherches technologiques et méthodologiques sont en cours ou devraient être entreprises ; elles concernent :

- des recherches sur les indicateurs les plus pertinents en matière d'audience des médias (mesure d'audience et qualification des internautes, ...),
- des tests sur le déroulement du questionnaire ou sur la formulation des questions (effet d'ordre de la présentation des titres selon les familles, formulation des questions sur la Lecture Dernière Période ou sur la Lecture d'un Numéro Moyen, ...),
- des recherches sur la comparaison d'audience inter-médias,
- des recherches sur la qualité des échantillons (amélioration des taux de

[3] Enquête comportant 15 000 interviews d' individus de 15 ans et plus, réalisée en face à face avec tirage aléatoire des îlots et méthode des quotas. La mesure d'audience de la Presse Quotidienne et de la Presse Hebdomadaire Régionale est réalisée dans une étude distincte de 20 000 individus interviewés par téléphone.

participation et des taux de réponse, biais dus aux non-réponses, ...),

- de nouvelles technologies de recueil de l'information dont l'objectif est essentiellement de diminuer l'effort demandé aux panélistes (audimètres, radiomètres, logiciels cartographiques pour la reconstitution des déplacements, enquêtes assistées par ordinateur, ...),
- des recherches sur les méthodes d'injection et de fusion d'enquêtes.

Bibliographie

Affichage Modèle 89, Centre d'Etude de Supports des Publicités et Chambre Syndicale Française de l'Affichage, 1989.

Advertising Research Foundation *Guidelines for Newspaper Audience Studies*, Advertising Research Foundation.

Bahu-Leyser D., Chavenon H., Durand J. (1990), *Audience des médias*, Guide France-Europe, Eyrolles.

CESP (1999), *Le Médiagraphe*

Dussaix A.M. (1999), *La mesure d'audience des médias*, Journal de la Société Française de Statistique, à paraître.

European Broadcasting Union (1997):

- *Guidelines for the setting up and operating of TV audience measurement peoplemeter systems,*
- *Non-domestic viewing holiday homes international TV channels,*
- *Towards harmonization* of radio audience measurement systems.

Gane R., "Radio audience measurement in the future", dans "2nd Radio Research Symposium", Warsaw, July 1997, ESOMAR Publication Series, vol. 214.

Guide des mesures d'audience 1996, Centre d'Etude des Supports de Publicité.

L'audience et les médias (1989), Institut de Recherches et d'Etudes Publicitaires, Les Editions d'Organisation.

Readership measurement in Europe (1996), 1996 Report on newspaper and magazine readership measurement in Europe, ESOMAR, Amsterdam.

ANALYSE DES DONNEES D'ENQUETES, DATA MINING ET TEXT MINING

Ludovic Lebart,
Directeur de recherche au Centre National de Recherche Scientifique

Ecole Nationale Supérieure de Télécommunication
46 Rue Barrault - 75013 PARIS - France
lebart@eco.enst.fr

Les techniques de traitement des données d'enquêtes ont été profondément modifiées par l'analyse des données (principalement ici: analyse en composantes principales, analyse des correspondances simples et multiples, classification automatique) qui intervient, dans une phase préliminaire, pour apprécier la qualité de l'information, synthétiser cette information, et orienter la suite des traitements.

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

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

1. Visualisation de données numériques

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

a) Méthodes factorielles

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

Mais la géométrie des nuages de points et les calculs de proximités ou de distances

qui en découlent diffèrent selon la nature des lignes et des colonnes du tableau analysé.

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

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

On rappelle brièvement ici trois techniques fondamentales :

- l'analyse en composantes principales (ACP) (Hotelling, 1933) s'applique aux tableaux de type "variables-individus", dont les colonnes sont des variables à valeurs numériques continues et dont les lignes sont des individus, des observations, des objets, etc. Les proximités entre variables s'interprètent en termes de corrélation ; les proximités entre individus s'interprètent en termes de similitudes des valeurs observées. Elle peut donner lieu à de nombreuses variantes en s'appliquant par exemple à un tableau de rangs (diagonalisation de la matrice de corrélation des rangs de Spearman), ou encore après l'élimination de l'effet de certaines variables (analyses locales ou partielles).
- l'analyse des correspondances (AC) (Benzécri, 1973) s'applique aux tableaux de contingences, c'est-à-dire aux tableaux de comptages obtenus par le croisement de deux variables nominales. Ces tableaux ont la particularité de faire jouer un rôle identique aux lignes et aux colonnes. L'analyse fournit des représentations des associations entre lignes et colonnes de ces tableaux, fondées sur une distance entre profils (qui sont des vecteurs de fréquences conditionnelles) désignée sous le nom de distance du Chi-deux .
- l'analyse des correspondances multiples (ACM) est une extension du domaine d'application de l'analyse des correspondances, avec cependant des procédures de calcul et des règles d'interprétation spécifiques. Son champ d'application est considérable. Elle est particulièrement adaptée à la description de grands tableaux de variables nominales dont les fichiers d'enquêtes socio-économiques ou médicales constituent des exemples privilégiés. Les lignes de ces tableaux sont en général des individus ou observations (il peut en exister plusieurs milliers); les colonnes sont des modalités de variables nominales, le plus souvent des modalités de réponses à des questions (cf. Lebart et al., 1995; Saporta, 1990).

b) Méthodes de classification

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

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

fonctionner avec plusieurs critères d'agrégation. Les deux principaux critères concernent la technique "du saut minimal", équivalente, d'un certain point de vue, à la recherche de l'arbre de longueur minimale (procédure classique en recherche opérationnelle) et la technique d'agrégation "selon la variance" ou de Ward, intéressante par la compatibilité de ses résultats avec certaines analyses factorielles.

Ces techniques présentent des avantages différents et peuvent être utilisées conjointement. Il est ainsi possible d'envisager une stratégie de classification basée sur un algorithme mixte, particulièrement adapté au partitionnement d'ensembles de données comprenant des milliers d'individus à classer. Un des avantages des méthodes de classification est de donner lieu à des éléments (les classes) souvent plus faciles à décrire automatiquement que les axes factoriels.

Enfin, la pratique montre que l'utilisateur a intérêt à utiliser de façon conjointe les méthodes factorielles et les méthodes de classification (cf. section 2.3 ci-dessous).

c) Problèmes statistiques et numériques liés à l'échelle des données

L'utilisation des méthodes précédentes dans le cadre du Data Mining, caractérisé par des dimensions considérables des tableaux à analyser, conduit à mettre en oeuvre des algorithmes spécifiques.

Dans le cas des méthodes factorielles, il s'agit, pour l'essentiel, de pouvoir calculer les matrices à diagonaliser (matrices de corrélations par exemple dans le cas de l'ACP) sans trop d'occupation mémoire et avec une précision convenable (ce qui n'est pas évident si les calculs se font sur des centaines de milliers d'observation), puis de diagonaliser ces matrices. Le problème est moins crucial avec les extensions de mémoire permises par les ordinateurs personnels actuels (on peut facilement diagonaliser une matrice 3000x3000 sans problème). Des algorithmes d'approximation stochastique (Benzecri, 1969) ou plus généralement des méthodes de gradients stochastiques permettent de procéder à des calculs en ligne, sans incorporer de tableaux volumineux en mémoire.

Ces méthodes d'un emploi très général (mais assez peu efficaces en terme de temps calcul et de qualité de la convergence) sont utilisées pour estimer les paramètres des réseaux de neurones (cf. section 2.5 ci-dessous).

Pour les méthodes de classification, ce sont les méthodes de type k-means et leurs méthodes dérivées qui permettent d'obtenir des partitions pratiquement sans limites sur le nombre d'individus à classer.

2. Le traitement des données d'enquêtes

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

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

Les techniques d'analyse des données (analyses descriptives multidimensionnelles) présentées en section 1 modifient profondément les premières phases du traitement des

données d'enquête. Elles vont en fait bouleverser l'enchaînement des tâches, et définir une méthodologie nouvelle.

a) Les étapes du traitement

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

1) *Descriptions élémentaires* (tri-à-plat, histogrammes, calculs de statistiques élémentaires, moyennes, écarts-types, valeurs extrêmes, quantiles). Retour éventuel aux données de base pour une nouvelle saisie partielle ou pour des corrections.
2) *Épreuves de cohérence globale* ; Épreuves d'hypothèses larges (par hypothèses larges, on entend : hypothèses générales permises par les nouveaux outils de description). Structuration des données, typologies, sélection de tableaux croisés.
3) *Épreuves d'hypothèses classiques* (tests statistiques usuels, régression, discrimination, analyses de la variance, modèles log-linéaires...).
4) *Conclusions* : Critique de l'information de base : lacunes dans le choix des variables, déséquilibre de l'échantillon ou du champ d'observation, biais ou erreurs. Choix de modèles, énoncés des résultats, rejets d'hypothèses, suggestions de nouvelles hypothèses.

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

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

Notons que les règles d'interprétation des représentations obtenues à l'issue des techniques de réduction présentées en section 1 n'ont pas la simplicité de celles de la statistique descriptive élémentaire. L'interprétation des histogrammes, des "camemberts", des graphiques de séries chronologiques est intuitive, alors que dans le cas de l'analyse des correspondances, par exemple, il sera nécessaire de connaître des règles de lecture des résultats. Une formation et une expérience pratique s'avéreront nécessaires.

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

L'analyse des correspondances et l'analyse en composantes principales, permettent de trouver des sous-espaces de représentation des proximités entre profils ou entre vecteurs de description d'observations. Mais elles permettent aussi de positionner dans ce sous-espace des lignes ou des colonnes supplémentaires du tableau de données (Cazes, 1981).

On peut ainsi illustrer les plans factoriels par des informations n'ayant pas participé à la construction de ces plans, ce qui va avoir des conséquences très importantes au niveau de l'interprétation des résultats.

Les éléments ou variables servant à calculer les plans factoriels sont appelés

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

Cette dichotomie entre variables actives et variables illustratives est du même ordre que la distinction que l'on établit entre variables exogènes (explicatives) et endogènes (à expliquer) dans les modèles de régression multiple. D'un point de vue géométrique, les deux situations sont d'ailleurs très similaires. Les variables exogènes engendrent un sous-espace sur lequel seront projetées les variables endogènes. Les variables actives engendrent aussi un sous-espace, que l'on va réduire pour le visualiser, et c'est sur cet espace réduit que l'on projette les variables illustratives.

c) Complémentarité de la classification

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

1. Les visualisations sont limitées à deux ou en général à très peu de dimensions, alors que le nombre d'axes significatifs peut souvent atteindre 8 ou 10, pour fixer les idées.
2. Ces visualisations peuvent inclure des centaines de points, et donner lieu à des graphiques chargés ou illisibles. Il faut donc à ce stade faire appel de nouveau aux capacités de gestion et de calcul de l'ordinateur pour compléter, alléger et clarifier la présentation des résultats.

L'utilisation conjointe de la classification automatique et des analyses précédentes permet de remédier à ces lacunes.

Lorsqu'il y a trop de points sur un graphique, il paraît utile de procéder à des regroupements en familles homogènes. Mais les algorithmes utilisés pour ces regroupements fonctionnent de la même façon, que les points soient situés dans un espace à deux ou à 100 dimensions.

Autrement dit, l'opération va présenter un double intérêt : allégement des sorties graphiques d'une part, prise en compte de la dimension réelle du nuage de points d'autre part.

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

d) Sélection raisonnée des tableaux croisés et noyaux factuels

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

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

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

On désigne par *noyaux factuels* des groupes d'individus les plus homogènes possibles vis-à-vis de leurs caractéristiques de base. On aimerait en effet croiser des caractéristiques telles que l'âge, le sexe, la profession, le niveau d'instruction, de façon à étudier des groupes d'individus tout à fait comparables entre eux du point de vue de leur situation objective (réaliser, dans la mesure du possible, le toutes choses égales par ailleurs). Mais de tels croisements conduisent vite à des milliers de modalités, dont on ne sait que faire lorsqu'on étudie un échantillon lui-même de l'ordre de quelques milliers d'individus. De plus, les croisements ne tiennent pas compte du réseau d'interrelations existant entre ces caractéristiques : certaines sont évidentes (il n'y a pas de "moins de 30 ans" retraités), d'autres sont également connues a priori, avec cependant des exceptions (il y a peu d'étudiants veufs), d'autres enfin ont un caractère plus statistique (il y a plus de femmes dans la catégorie "plus de 60 ans"). Une classification des individus décrits par la batterie active des caractéristiques de base va permettre de regrouper les individus ayant, dans l'échantillon, le maximum de caractéristiques en commun. En pratique, elle fournira des regroupements opératoires en une vingtaine de classes pour un échantillon de l'ordre de 2 000 individus. Le tableau croisant une des variables nominales de l'enquête avec la partition en noyaux factuels résume pratiquement tous les tableaux obtenus en croisant cette même variable avec chacune des caractéristiques de base. De plus, certaines interactions indécelables à partir de ces tableaux binaires peuvent être détectées.

e) Itération de traitements. Articulation description-inférence.

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

Les opérations de recodage font partie d'un processus itératif qui converge vers une connaissance et une assimilation optimale de l'information de base. La panoplie du statisticien contient des modèles, permettant, à partir de variables quelconques, de prévoir une variable numérique (régression, analyse de la variance et de la covariance), une variable nominale (analyse discriminante, régression logistique: cf.: Bardos, 1989; Celeux et Nakache, 1994; Hand, 1997), d'étudier les associations dans les tables de contingence (modèles d'association, modèles log-linéaires). La régression et l'analyse discriminante par arbre (Breiman et al., 1984), qui améliore les méthodes classiques de segmentation utilisées en marketing, est une des méthodes prédictive les plus utilisées dans le cadre du Data Mining. Enfin les réseaux de neurones (Hérault et Jutten, 1994; Thiria et al., 1997) constituent des modèles souples et non-linéaires qui généralisent la plupart des méthodes précitées. Leur fonctionnement en tant que boite noire, la difficulté d'interprétation des paramètres, les problèmes de convergence numérique font que ces méthodes ne se substituent pas totalement aux méthodes statistiques plus classiques.

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

3. Visualisation de données textuelles

Les analyses statistiques de textes peuvent intervenir à deux niveaux dans un contexte industriel et commercial: au niveau du traitement des lettres de réclamations, des cahiers de doléances ou de suggestion, au niveau de questions ouvertes dans des enquêtes postales ou téléphoniques. Les questions ouvertes les plus simples et les plus fréquentes sont d'une part la question "pourquoi" posée après une question fermée, et d'autre part les questions du type "autre, préciser", comme item de réponse complémentaire à une question fermée. Le traitement proposé va produire de façon automatique des mots caractéristiques et des réponses caractéristiques pour diverses catégories de répondants.

a) Questions ouvertes dans les enquêtes

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

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

b) Les traitements statistiques de textes

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

- Les applications à des textes littéraires (attributions d'auteurs, datation, par exemple) qui cherchent à saisir des caractéristiques de forme et de style à partir des distributions statistiques de vocabulaire, d'indices ou de ratios, ou encore à partir de corpus partiels de mots-outil (articles, conjonction, etc.). Il s'agit de saisir les "invariants" d'un auteur ou d'une époque, dissimulés ou peu apparents, à des fins historiques, littéraires, dans le cadre d'études que l'on désigne sous le nom de stylométrie (cf. par exemple Holmes, 1985, pour une revue de ces travaux).

- D'autre part les applications réalisées en recherche documentaire (Salton, 1988), en codification automatique, dans le traitement des réponses à des questions ouvertes, qui s'intéressent principalement au contenu, au sens des textes. Cependant, lors du traitement statistique de réponses à des questions ouvertes, ou lors des analyses d'entretiens, le socio-linguiste peut être aussi intéressé par la forme, par les connotations véhiculées par exemple par certains synonymes, certaines tournures (cf. par exemple : Achard, 1993). Les méthodes d'analyses de réponses libres dans les enquêtes relèvent de cette seconde famille d'application (Lebart et Salem, 1994).

c) Les unités statistiques découpées dans les textes

Les formes graphiques

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

"Mots-outil", parties du discours

La notion de mot-outil (appelés encore mots vide en documentation, ou mots grammaticaux) ne se prête à aucune formalisation satisfaisante. Dans le tableau précédent, de, des, le, la, les, peuvent être retenus comme mots-outils, bien que le, la, les, aient plusieurs statut grammaticaux possibles. La recherche d'attribution d'auteur a longtemps privilégié l'étude de ce type d'unité, posant que leur emploi, moins maîtrisé lors de la rédaction du texte et plus indépendants du contenu, pouvait constituer une marque d'auteur privilégiée.

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

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

Les unités lemmatisées

Un autre type de traitement préliminaire du texte consiste à procéder à une lemmatisation. Cette opération, difficile à réaliser de façon entièrement automatique, consiste à remplacer les formes par l'entrée du dictionnaire correspondant (infinitif pour les verbes, masculin singulier pour les adjectifs, formes non élidées à la place des formes élidées, etc.). Elle est parfois complétée par la suppression de certains mots-outils (articles, conjonctions, etc., cf. par exemple Reinert, 1986). En documentation automatique, cela permet de travailler avec un nombre restreint de mots-clé dont les occurrences sont fréquentes. Une lemmatisation complète demande une analyse morpho-syntaxique approfondie, et ne peut être entièrement automatique (cf. Charniak, 1993).

En traitement de questions ouvertes, cette opération n'est pas toujours souhaitable a priori car elle détruit certaines locutions. En revanche, elle peut intervenir comme complément, car elle fournit un point de vue différent de celui fourni par une analyse entièrement automatique sur les formes graphiques du texte. Dans le cas d'entretiens non directifs peu nombreux, la lemmatisation permet de travailler avec des seuils de fréquences de mots plus élevés que ceux nécessités par l'analyse des formes graphiques.

d) Les analyses statistiques; les trois outils de base.

Une numérisation préliminaire (qui est aussi une compression) consiste à affecter à chaque nouvelle forme graphique un numéro d'ordre qui sera associé à toutes les occurrences de cette même forme. Ces numéros seront stockés dans un dictionnaire de formes, ou vocabulaire, propre à chaque exploitation.

Les trois outils de base sont l'analyse des correspondances des tableaux lexicaux, les sélections de formes caractéristiques, les sélection de réponses modales.

Analyse des correspondances des tableaux lexicaux

Les analyses des correspondances (cf. section 1) peuvent décrire les tables de contingence croisant les réponses et les formes graphiques, ou des groupes de réponses (par exemple regroupement selon le niveau d'instruction des répondants) et les formes graphiques. Elles permettent de visualiser les associations entre mots (formes) et groupes ou modalités. Ainsi, une visualisation des proximités entre mots et catégories socioprofessionnelles pourra aider la lecture des réponses de chacune de ces catégories.

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

De même, la présence de plusieurs flexions d'un même verbe constitue un outil de validation. Ainsi, une analyse appliquée à des données d'enquêtes (cf. Lebart, 1982) a pu montrer une opposition entre les formes (devoir, doit, doivent) d'une part, et les formes (pouvoir, peuvent, peut) d'autre part. L'existence de plusieurs formes relatives à un même lemme constitue une véritable épreuve de validité des résultats.

Formes ou segments caractéristiques (ou spécificités)

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

Les sélections des réponses modales

Pour un groupe d'individus donné, et donc pour le regroupement de réponses correspondant, les réponses modales (ou encore phrases caractéristiques, ou documents-type, selon les domaines d'application) sont des réponses originales du corpus de base, ayant la propriété de caractériser au mieux la classe.

On peut, pour chaque regroupement, calculer la distances du profil lexical d'un individu au profil lexical moyen du groupement. On peut ensuite classer les distances par ordre croissant, et donc sélectionner les réponses les plus représentatives au sens du profil lexical, qui correspondront aux plus petites distances. On obtient ainsi une sorte de résumé des réponses de chaque regroupement, formé de réponses originales.

e) Stratégie de traitement

On a vu qu'il était souvent nécessaire de regrouper les réponses pour pouvoir procéder à des analyses de type statistique. Les profils lexicaux d'agrégats de réponses ont plus de régularité et de signification que ceux des réponses isolées. Ce regroupement a priori peut être réalisé à partir des variables disponibles, retenues en

fonction de certaines hypothèses. Mais ceci suppose une bonne connaissance préalable du phénomène étudié, situation qui n'est en général pas réalisée dans les études dites exploratoires.

Regroupement par noyaux factuels

La technique dite des "noyaux factuels" déjà évoquée en section 2 va permettre de donner des éléments de réponse à ce problème. Étant donnée une liste de descripteurs ou de variables caractérisant les individus, le problème est de regrouper les individus en groupes les plus homogènes possibles vis-à-vis de ces caractéristiques, sans en privilégier certaines a priori. La partition obtenue est une sorte de "partition moyenne" qui résume les principales combinaisons de situations observables dans l'échantillon, et qui permet donc de procéder à des regroupements de réponses textuelles les moins arbitraires possibles.

Analyses directes sans regroupement

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

Il est donc possible d'illustrer ces typologies par les caractéristiques des individus interrogés qui auront le statut de variables supplémentaires ou illustratives. Ce traitement direct des réponses pourra conduire à la réalisation d'un post-codage partiellement automatisé.

f) Conclusions sur les analyses de textes

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

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

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

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

4. Problèmes de qualité d'information

Les visualisations de variables numériques ou textuelles qui viennent d'être évoquées permettent de prendre en compte certaines déficiences de l'information de base (non-réponses par exemple), ainsi que des variables de contrôles liées à la qualité du recueil de l'information de base (cf. ASU, 1993).

a) Conjectures sur les non-réponses

Les non-réponses sont des modalités comme les autres, qui peuvent être positionnées dans les espaces factoriels des thèmes, comme dans les espaces de la structure de base. Le traitement des non-réponses qui se prête mal aux tests statistiques usuels reçoit ici une importante contribution, dans la mesure où l'on peut étudier le contexte de ces refus ou lacunes, soit en termes de caractéristiques des répondants, soit à partir des réponses effectives à d'autres thèmes.

b) Le positionnement des variables techniques

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

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

c) Les questions ouvertes

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

Conclusion

Finalement, dans le traitement des données d'enquêtes comme dans le data mining ou le text mining en général, l'approche globale et exploratoire est un élément important d'une démarche qualité. En effet, détecter des patterns, c'est évidemment dans une première phase détecter des anomalies, des incohérences, des points aberrants (outliers), des hétérogénéités inattendues. Dans le cas des données transactionnelles, ces résultats peuvent donner de précieuses informations sur les principes et la mise en oeuvre effective du système d'information de gestion de l'entreprise. Dans le cas des données d'enquêtes, particulièrement en présence de questions ouvertes, cela peut amener non seulement une critique de la réalisation pratique de l'enquête et du recueil de l'information sur le terrain, mais cela peut dans certains cas aller jusqu'à une remise en cause de la conception de l'enquête et de son questionnaire.

Références bibliographiques

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

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

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

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

Belson W.A., Duncan J.A., (1962) - A Comparison of the check-list and the open

response questioning system, Applied Statistics n°2, pp. 120-132.

Benzécri J.-P. (1969b) - Approximation stochastique dans une algèbre normée non commutative. Bull. Soc. Math. France, 97, p 225-241.

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

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

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

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

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

Deroo M., Dussaix A.-M. (1980) - Pratique et analyse des enquêtes par sondage. P.U.F., Paris.

Diday E. (1971) - La méthode des nuées dynamiques. Revue Statist. Appl. 19, n° 2, p 19-34.

Escofier B., Pagès J. (1988) - Analyses factorielles simples et multiples. Dunod, Paris.

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

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

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

Hand D. J. (1998) - Data mining: Statistics and more? The American Statistician (May issue).

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

Holmes D.I., 1985, The analysis of literary style - A Review, J.R.Statist.Soc., 148, Part 4, pp. 328-341.

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

Lafon P., 1980, 'Sur la variabilité de la fréquence des formes dans un corpus', Mots, 1, pp.127-65.

Lebart L., 1982, L'Analyse statistique des réponses libres dans les enquêtes socio-économiques, Consommation, n°1, Dunod, pp. 39-62.

Lebart L., Morineau A., Piron M. (1995) - Statistique Exploratoire Multidimensionnelle. Dunod, Paris.

Lebart L., Salem A. (1994) - Statistique textuelle. Dunod, Paris.

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

Reinert M., 1986, Un Logiciel d'analyse lexicale . Les cahiers de l'analyse des données, 4, Dunod, pp. 471-484.

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

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

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

MODELISATION SPATIALE DU TRAFIC TELEPHONIQUE ET SIMULATIONS

Jean BARBE
Cour des comptes européenne - Luxembourg
e-mail : jean.barbe@eca.eu.int

Jean-Sébastien ROY
Caisse Nationale d'Assurance Maladie des Travailleurs Sociaux - Paris

Pourquoi modéliser le trafic téléphonique à un niveau fin ?

Un des facteurs fondamentaux du coût d'un service téléphonique et de son organisation concurrentielle est l'organisation géographique du trafic : (zone A, zone B, trafic entre A et B).

De nombreux modèles de coûts (utilisés en interne par l'Opérateur de Télécommunications ou en externe, par exemple par l'Autorité de Régulation) utilisent des approximations (comme la densité moyenne).

La rareté des modélisations de trafic[1] sur des zones fines tient principalemet à :

- ces données proviennent d'entités techniques gestionnaires du "Réseau" éloignées d'entités financières ou économiques[2],
- pour prendre en compte la dimension spatiale, ces données sont à utiliser à des niveaux très fins, à ces niveaux là, il est difficile de trouver une information démographique et économique,
- les modèles doivent prendre en compte des micro-phénomènes complexes (les frontières par exemple).

Ces modèles sont pourtant essentiels pour quantifier la prospective économique et stratégique du secteur :

- comment l'évolution de la répartition sur le territoire de la démographie (population et activité économique) va-t-elle modifier les grands axes de trafic téléphonique ?
- va-t-on vers l'apparition de nouveaux axes interurbains, le renforcement d'axes de grandes distances géographiques, ou la dilution du trafic sur le territoire ?
- in fine comment évolueront les coûts et à réglementation et jeux d'acteurs donnés, l'intensité concurrentielle ?

Les différentes phases de l'étude

L'étude suit les 5 phases suivantes, de la collecte des données à la production d'outils à la décision :

1. Collecte des données : 5 sources
2. Data management
   - travail des données pour les rendres "fusionnables"
   - fusion des données

[1] Par contre les modèles macro-économiques de demande de trafics sortants sont courants à de nombreux niveaux.

[2] Dans le cas de France Telecom, Branche Réseaux / Branches Développement, Grand Public ou Entreprise.

3. Analyse des données
   - contrôle de qualité
   - prise en main de la base
4. Modélisation
   - prise en compte des contraintes
   - spécification
   - calcul des différentes distances
   - estimation, tests comparatifs et analyse critique des différents modèles
   - scénarios de simulation
5. Outil d'aide à la décision
   - par scénario : variation du trafic en erlang[3] par longueur d'axe, par importance en erlang de l'axe
   - piste d'étude pour l'amélioration des modèles et de l'outil d'aide à la décision

Collecte des données

La collecte des données s'organise selon 5 domaines distincts :

1. Données de trafic
   - Nombre d'erlangs, entre les ZAA[4], y compris l'international, pour le jour et le soir.
2. Données de tarification
   - Nombre de secondes par unité entre chaque CT[5].
3. Données cartographiques
   - Position et surface de chaque commune (données MAPINFO).
4. Données de parc
   - Nombre de lignes selon la segmentation RPXYZ[6], par ZAA
5. Données économiques :
   - Données du fichier SIRENE (Nombre d'entreprises et d'employés, selon le secteur d'activité), au niveau de chaque commune.
   - Données du recensement (Population française et étrangère), au niveau de chaque commune.

Data Management

Le premier travail sur les données consiste à les rendres "fusionnables" :

- pour chaque source, une table de passage a été créée pour aboutir à une description en terme de ZAA,
- cependant la dimension commune, en général beaucoup plus fine (36 000 communes pour 400 ZAA) a été utilisée de manière intermédaire dans la modélisation pour le calcul des distances.

---

[3] L'erlang est l'unité de trafic et correspond à l'occupation d'une ligne pendant une heure.

[4] ZAA : zone d'autonomie d'acheminement. Au nombre de 400 environ en France, cette notion est seulement technique, mais il s'agit de la plus petite unité disponible.

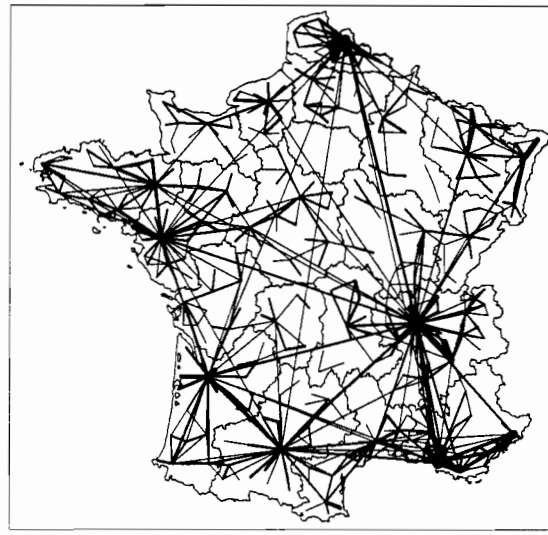
[5] CT : circonscription tarifaire, notion technique traduisible en terme de ZAA.

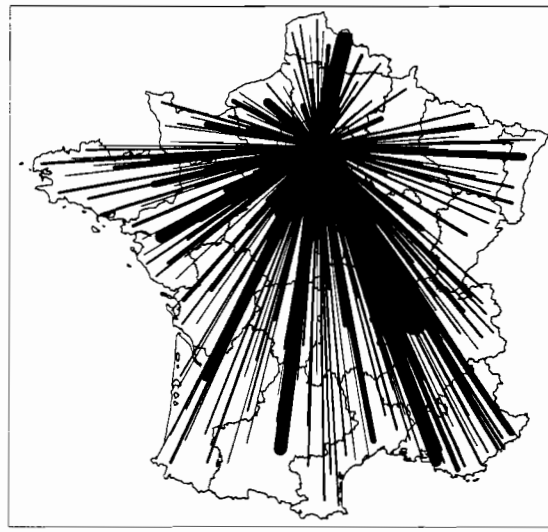
[6] RPXYZ : notions commerciales (respectivement résidentiels, professionnels et entreprises de plus en plus consommatrices de services de télécommunications).

La fusion des données a été réalisée à l'aide du logiciel SAS, MAPINFO ayant servi à des calculs de distances.

Le premier résultat : un trafic très concentré au départ et depuis Paris

- Trafic mesuré (excepté avec Paris, et avec Paris) (1000 liens les plus importants)





Analyse des données

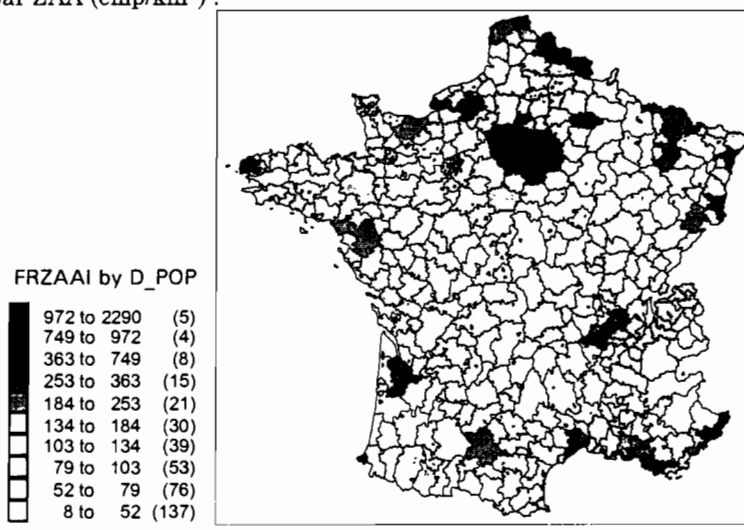
Les données tarifaires sont redondantes, une fois pris en compte la distance, elles ne seront donc pas utilisées dans la modélisation. Les données de parc sont écartées pour ne pas à être prévues au moment des simulations.

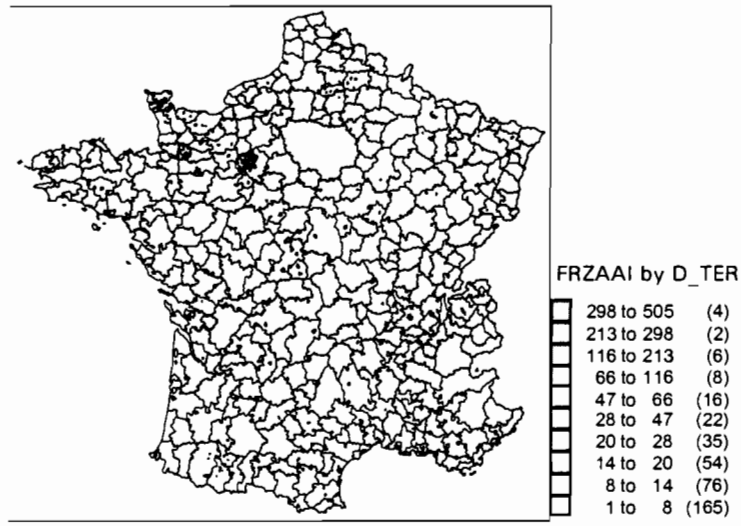
Certaines communes sont partagées entre plusieurs ZAA, d'autres, par contre comprennent plusieurs ZAA, on a du faire dans le premier cas des affectations au prorata, dans le second cas sommer les erlangs (ce qui suppose le parallélisme des courbes

de charge).

Les quatre ZAA les plus importantes en terme de trafic sont l'International, Paris, Lyon, Marseille et Lille. Les cinq premières destinations/origines accumulent à elles seules plus de 40% du trafic. Un quart de celui-ci est effectué avec la ZAA de la région parisienne.

Densité de population par ZAA (hab/km^2) et Densité d'emplois dans le secteur tertiaire par ZAA (emp/km^2) :





Modélisation : contraintes de modélisation et spécifications

Le modèle doit vérifier les contraintes suivantes :

- Le modèle doit être indépendant du maillage. (⇒ bilinéarité du numérateur à dénominateur quasi-constant, cf. ci-dessous)
- Le modèle doit être symétrique vis-à-vis des deux ZAA concernées, le trafic s'étant avéré quasi-symétrique. Il est néanmoins possible de se passer de cette

contrainte de symétrie.

- Le modèle ne doit utiliser que des paramètres susceptibles d'intervenir dans une simulation.

On en déduit les impacts suivants sur les spécifications :

- On note A et B deux ZAA, P_A et P_B leurs populations respectives, T_{AB} le trafic entre A et B (bidirectionnel) et D_{AB} une distance (cf. ci-après) entre les ZAA. Les paramètres sont notés k et p.
- Les modèles testés sont alors sensiblement, pour le jour et le soir, des combinaisons de modèles de la forme : $T_{AB} = \frac{kP_A P_B}{D_{AB}^p}$, faisant intervenir des populations différentes : habitants, étrangers, employés, employés selon le secteur d'activité, nombre de lignes RPXYZ, le trafic vers l'international, etc.
- Des modèles différents sont calculés selon un nombre variable de classes de distances, elles aussi paramètres du modèle.
- Le choix des modèles s'est fait en se basant sur un arbitrage simplicité / qualité.

Les trois chapitres suivants listent les points clefs de la modélisation et l'apport par rapport à un modèle gravitaire classique :

- Distances en "population"
- Modélisation infra ZAA au niveau de la commune,
- Modélisation hiérarchique : selon la longueur d'un lien, la spécification retenue est adaptée.

Modélisation : distances en population

Les modèles classiques de trafic de type *gravitaire* [$T_{AB} = \frac{kP_A P_B}{D_{AB}^p}$] ont été testés. Cette distance explique significativement le trafic avec une valeur optimale de p voisine de 1. Dans la suite on fixera p égal à 1.

Une nouvelle distance faisant intervenir, outre la distance à vol d'oiseau, la population ou les emplois a été calculée : l'hypothèse sous-jacente est de considérer que la propension d'une personne à en appeler une autre est inversement proportionnelle au nombre de personnes situées à une distance inférieure de celle de son correspondant.[7]

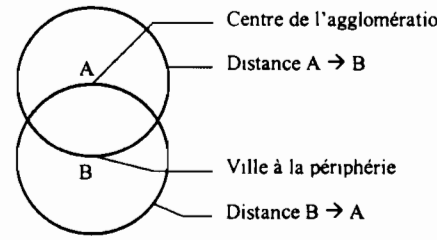
Une telle distance permet donc d'inclure dans le modèle des effets de densité, de répartitions des populations, selon que la population se concentre autour d'une ville, s'aligne le long d'une vallée, etc.

Cette distance est donc calculée comme l'ensemble de la population située dans un disque de centre l'appelant et de rayon égal à la distance entre l'appelant et l'appelé.

Cette distance permet *si nécessaire*, aussi de s'affranchir de la contrainte de symétrie.

[7]Par exemple, en rase campagne, peu importe que le village le plus proche soit à 20, 30 ou 50 km : pour joindre quelqu'un, il faut téléphoner à ce village. Ainsi, la distance ne compte que si il y a des personnes à joindre situées plus près.

- Une forme du type $T_{A\to B} = \frac{kP_A P_B}{D_{A\to B}}$ permet de rendre compte de l'asymétrie de trafic qu'il est possible d'observer entre le centre et la périphérie d'une agglomération.
- La population, représentée en gris (voir dessin), est bien plus importante dans le cercle de centre A (le centre de la ville), (cette population est la distance A→ B), que la population située dans le cercle de centre B (un quartier périphérique) (distance B→ A).



Modélisation : modélisation infra ZAA (niveau commune)

Les ZAA ne sont pas homogènes vis-à-vis de la répartition de la population : elles regroupent des zones désertiques et des zones excessivement denses (la Forêt de Fontainebleau et la Défense sont toutes deux situées dans la ZAA parisienne).

Pour éviter cette perte d'information par agrégation, on peut calculer les distances au niveau communale, puis agréger la distance au niveau des ZAA, seul niveau où le trafic est disponible, donc la modélisation possible.

- Si I est une commune de A, J une commune de B, et $D_{I,J}$ la distance en population entre I et J, le trafic entre I et J suit le modèle : $T_{I,J} = \frac{kP_I P_J}{D_{I,J}}$.

  Sous l'hypothèse que des communes proches ont des courbes de charge semblables, on additionne les trafics en Erlang :

  $$\sum_{I\in A, J\in B} T_{I,J} = k \sum_{I\in A, J\in B} \frac{P_I P_J}{D_{I,J}} \quad \text{soit} \quad T_{A,B} = k \sum_{I\in A, J\in B} \frac{P_I P_J}{D_{I,J}}$$

- Cette distance est longue à calculer, (balayage sur AxB).
- Soit au total 49 mille milliards de communes à examiner. On réduit un peu ce nombre via un ordonnancement adéquat des communes.

Modélisation hiérarchique

Trois facteurs sont susceptibles d'engendrer des non-linéarités :

- Paris / la province : la place à part de Paris,
- La distance : l'affaiblissement de l'effet de la distance, plus radical que la simple décroissance d'une hyperbole et moins monotone,
- Le jour / le soir : les trafics des entreprises et des résidentiels suivent a priori des lois différentes. S'ils sont mélangés, c'est dans des proportions différentes, le jour et le soir.

Les ruptures Paris / Province et Jour / Soir sont faciles à tester.

Pour les ruptures dues à la distance, des itérations sur les distances limite assorties de statistiques globales sur les modèles induits, ont permis de choisir de manière endogène les seuils optimaux.

Comme pour les ruptures, le choix comme variable explicative, d'un type de population (habitants, effectifs du secteur tertiaire, ou autres), la forme de la distance

et la manière avec laquelle elle intervient sont endogènes, i.e. le résultat d'optimisations visant à obtenir un modèle offrant un bon rapport qualité / complexité. Seules les spécifications discutées précédemment sont a priori, même si guidées par les contraintes techniques.

Modélisation : résultats

Les paramètres (tous significatifs) ont été estimés par les mco, les ruptures à 25 km, 70 km et 300 km ont été estimées par itérations.

- POP2 : produit des populations des deux ZAA
- ETR2 : produit des populations étrangères des deux ZAA
- DIST : distance entre les deux ZAA
- E_STER2 : produit des emplois du secteur tertiaire des deux ZAA
- PAR : somme des produits de populations divisé par la distance en population entre toutes les communes des deux ZAA.
- PAR3 : Idem, pour les emplois du secteur tertiaire.

Modèle Soir

Distance		Paramètres du modèle linéaire utilisé (et valeur du paramètre)
< 25		PAR : 0.0013446 ETR2 : $9.5769\ 10^{-7}$
25 à 70		PAR : 0.0010754
70 à 300		PAR : 0.00087277
> 300	Avec Paris	POP2 : $1.2777\ 10^{-10}$
> 300	Sans Paris	POP2/DIST : $3.1984\ 10^{-8}$

Modèle Jour

Distance		Paramètres du modèle linéaire utilisé (et valeur du paramètre)
< 25		PAR : -0.00037796 PAR3 : 0.0012017 ETR2 : $9.6575\ 10^{-7}$
25 à 70		PAR : 0.00077399 PAR3 : 0.00023039
70 à 300		PAR : 0.000086369 PAR3 : 0.0003735
> 300	Avec Paris	E_STER2 : $2.4902\ 10^{-9}$
> 300	Sans Paris	PAR3*DIST : $1.4204\ 10^{-6}$

Modélisation : qualité

On s'intéressera à la qualité intrinsèque du modèle, mais surtout aux améliorations successives, au fur et à mesure des amendements du modèle gravitaire élémentaire.

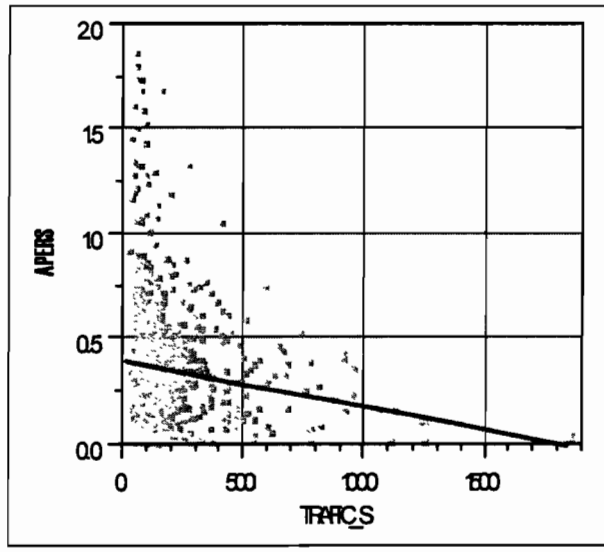
On peut tirer deux résumés des régressions :

1. Moins de 13 % d'erreur sur 25 % du trafic - moins de 28 % d'erreur sur 50 % du

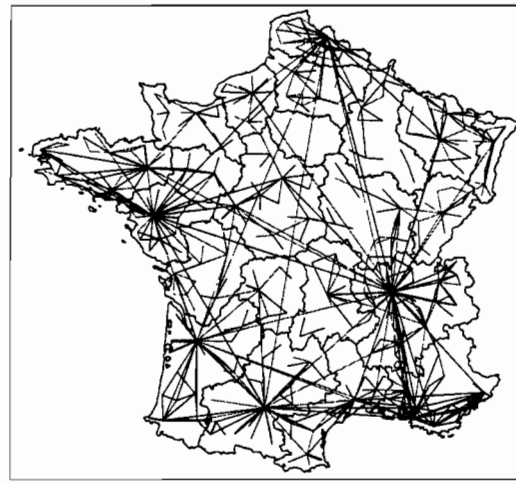
trafic - moins de 46 % d'erreur sur 75 % du trafic

2. Les liens les plus importants sont ceux pour lesquels l'erreur est la plus faible.

Erreur relative du modèle (APERS) selon la taille du lien (TRAFIC_S)



Différence entre le trafic mesuré et le modèle (Paris non représenté) (1000 liens les plus importants, en rouge : modèle > mesure, en bleu : modèle < mesure)



Modélisation : commentaires

Pour le modèle Jour :

- PAR3*DIST permet de faire intervenir la distance en population tout en supprimant l'effet de la distance à vol d'oiseau.

Modèle soir :

- POP2/DIST est utilisé à la place de PAR car la distance en population (habitants) n'a, pour les individus, plus grand sens au-delà d'une certaine distance.

Comparaison des deux modèles :

_ *Paris centre de la France :* dans les deux modèles, au delà de 300 kilomètres, la distance n'intervient plus pour les communications avec Paris. Ceci peu s'expliquer par le caractère cosmopolite de l'agglomération Parisienne, ainsi que sa position centrale en France.

_ *Jour Résidentiel et Business, Soir Résidentiel :* on observe que le trafic est déterminé le jour, à la fois par le nombre d'habitants et le nombre d'employés du tertiaire alors que seul le nombre d'habitants est utilisé dans la modélisation du trafic soir. Ceci montre que le trafic en soirée est majoritairement un trafic entre résidentiels. On remarque aussi que le trafic des entreprises est peu dépendant de la distance, au delà de 300 kilomètres.

Qualité des modèles et gains apportés par les différentes étapes (I)

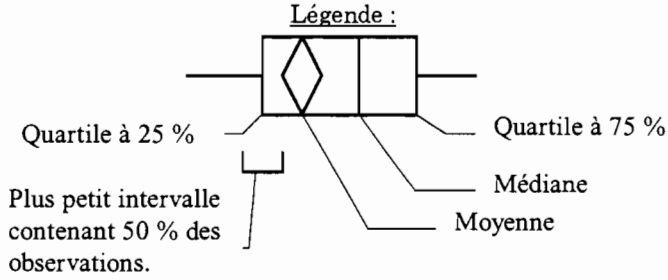
Les différents modèles qui ont été testés sont successivement :

1. $T_{AB} = kP_A P_B$,
2. $T_{AB} = k\frac{P_A P_B}{D_{AB}^p} k' E_A E_B$,
3. idem 2, avec ruptures selon la distance,
4. idem 3 avec distance en population au lieu de la distance à vol d'oiseau,
5. idem 4, avec une distance calculée au niveau de la commune et non de la ZAA.

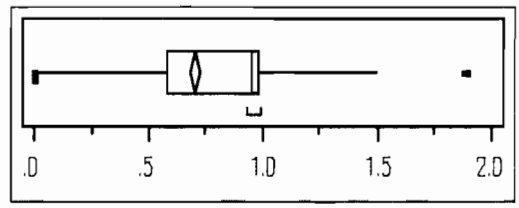
Le premier résultat est une amélioration à chaque étape.

_ Les graphes ci-dessous représentent la répartition des résidus relatifs (1=100% d'erreur) sur les 1000 liens sélectionnés, suivant le modèle utilisé.

_ On note que l'amélioration la plus importante est réalisée par l'utilisation de la distance en population. (Note : le gain dû à l'utilisation de la distance au niveau des communes est plus important pour le Jour que pour le Soir).

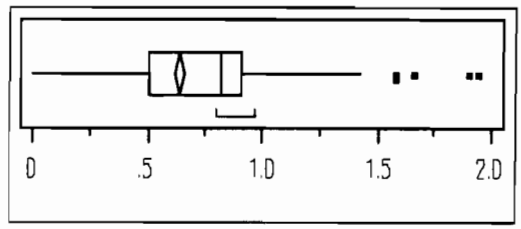


Modèle 1 (Pop2)



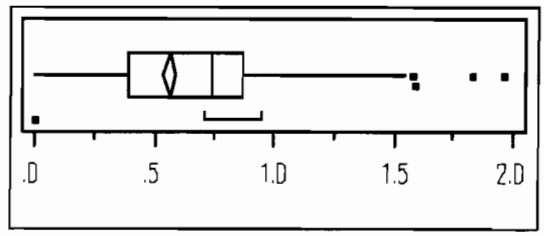
(Sur 50% du trafic on fait moins de 95% d'erreur)

Modèle 2 (Pop2/Dist, Etr2)



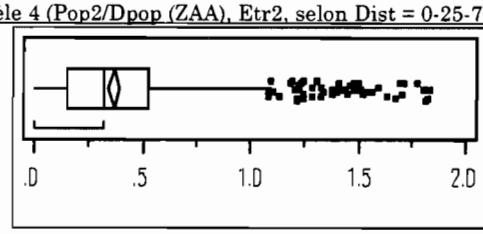
(Sur 50% du trafic on fait moins de 81% d'erreur)

Modèle 3 (Pop2/Dist, Etr2, selon Dist = 0-25-70-300-)



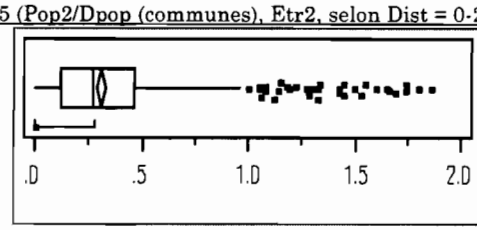
(Sur 50% du trafic on fait moins de 73% d'erreur)

Modèle 4 (Pop2/Dpop (ZAA), Etr2, selon Dist = 0-25-70-300-)



(Sur 50% du trafic on fait moins de 33% d'erreur)

Modèle 5 (Pop2/Dpop (communes), Etr2, selon Dist = 0-25-70-300)



(Sur 50% du trafic on fait moins de 28% d'erreur)

Simulations : concentration de la population et de l'activité

On arrive enfin au point crucial : l'utilisation de la modélisation. Deux simulations ont été réalisées : concentration urbaine et tertiarisation de l'économie.

A été estimé l'impact sur le trafic d'une concentration de la population et de l'activité dans les grandes villes ainsi que l'effet inverse.

La concentration a été modélisée comme une augmentation de 10 % de la population et de l'effectif employé dans les villes de moins de 50 000 habitants, redistribuée dans les villes de plus de 50 000 habitants, au prorata de leur population.

Le tableau suivant donne les résultats par importance des liens (en Erlang) et

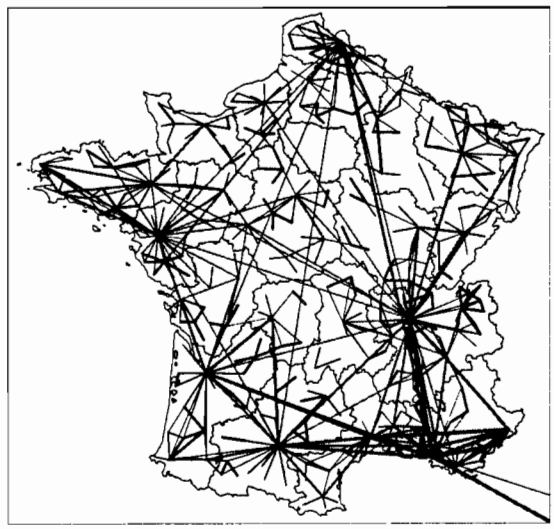
longueur (en km), qui sont les variables stratégiques primordiales en terme de coût et concurrence par exemple.

La carte représente les liens dont le trafic augmente en rouge, et ceux pour lesquels le trafic diminue en bleu.

On peut faire les remarques suivantes :

- les liens de moins de 25 km sont rares à cause de la taille des ZAA
- si la concentration provoque en moyenne une baisse de trafic *(entre ZAA)*, en diminuant le trafic entre les zones moins concentrées, elle provoque un accroissement de celui-ci sur les liens interurbains les plus importants. (sur la carte seuls les 1000 liens les plus importants hors Paris ont été représentés). La déconcentration provoque sensiblement les effets opposés à la concentration
- l'impact sur la concurrence est mixte : renforcement des liens interurbains *rentables* et du local (non pris en compte dans le modèle : trafic intra ZAA).

Distance (km)	*Taille des liens (erlangs)*	*% de liens concernés*	*% de trafic concerné*	*% de variation (erlangs, jour)*	*% de variation (erlangs, soir)*
Tous	Tous	100,00	100,0		
Tous	< 60	98,71	36,8		
Tous	> 60	1,41	63,2		
< 25	Tous	0,14	6,7		
25 à 70	Tous	2,21	31,9		
70 à 300	Tous	27,93	34,1		
> 300	Tous	69,71	27,3		
< 25	< 60	0,04	0,4		
25 à 70	< 60	1,48	6,2		
70 à 300	< 60	27,60	17,7		
> 300	< 60	69,46	12,5		
< 25	> 60	0,10	6,3		
25 à 70	> 60	0,73	25,7		
70 à 300	> 60	0,33	16,4		
> 300	> 60	0,25	14,8		



Simulations : tertiarisation de l'économie

La tertiarisation de l'économie est modélisée par un accroissement de 10 % de l'effectif du secteur tertiaire à emploi et population constante.

Une augmentation de 10 % de l'effectif du secteur tertiaire provoque ainsi une augmentation de plus de 20 % du trafic sur les liens interurbains les plus importants, les plus aisés à concurrencer.

Suit un tableau identique au précédent.

Distance (km)	*Taille des liens (erlangs)*	*% de liens concernés*	*% de trafic concerné*	*% de variation (erlangs, jour)*
Tous	Tous	100,00	100,0	
Tous	< 60	98,71	36,8	
Tous	> 60	1,41	63,2	
< 25	Tous	0,14	6,7	
25 à 70	Tous	2,21	31,9	
70 à 300	Tous	27,93	34,1	
> 300	Tous	69,71	27,3	
< 25	< 60	0,04	0,4	
25 à 70	< 60	1,48	6,2	
70 à 300	< 60	27,60	17,7	
> 300	< 60	69,46	12,5	
< 25	> 60	0,10	6,3	
25 à 70	> 60	0,73	25,7	
70 à 300	> 60	0,33	16,4	
> 300	> 60	0,25	14,8	

Conclusion

En conclusion présentons, les forces et les faiblesses du modèle, ainsi que les amendements possibles.

- Atouts du modèle :
  1. Ce modèle permet de quantifier la déformation de la demande de trafic téléphonique de point à point en fonction de scénarios économico-démographiques.
  2. Il est capable de détailler les résultats (i) géographiquement et en fonction de la (ii) taille des axes et de leur (iii) longueur.
  3. Ces trois aspects sont des facteurs explicatifs à la fois des *coûts* et de la *concurrence*.
  4. Cependant, si la modélisation marque un progrès par rapport à un modèle gravitaire simple, elle gagnerait à être développée selon certaines pistes découvertes à ce stade.
- Faiblesses :
  1. moindre robustesse des erlangs par rapport aux minutes (cf. le système d'information en amont de cette étude),
  2. modélisation simultanée du trafic inter ZAA et intra ZAA (cf. scénario de concentration)
- Développements :
  1. variables omises : l'analyse des résidus ouvre les pistes suivantes :
     - zones frontalière / côtière / intérieure,
     - zones touristiques
  2. meilleure prise en compte du trafic international :
     - éclatement de la ZAA fictive "international" en plusieurs groupes (Méditerranée - Union européenne - Amérique du nord...)

PROGRAMME DU 30 AVRIL 1999

8h30 : Séance d'ouverture

9h : ***Systèmes de diffusion de données financières en temps réel,***
par **Christian GOURIEROUX,**
professeur à l'Université Paris-Dauphine, directeur du Laboratoire Finance-Assurance du Centre de Recherche en Economie et STatistique.
9h30 : Débats

10h : Pause café

10h30 ***Les outils du micro marketing : Construction et exploitation d'un data warehouse pour le déploiement du marketing opérationnel,***
par **Jean-Michel GAUTIER,**
professeur à HEC, directeur général d'AXIS Conseil
11h : Débats

11h30: ***La mesure de l'audience des différents médias,***
par **Anne-Marie DUSSAIX,**
professeur à l'ESSEC, ancienne présidente de la SFdS.
12h : Débats

12h30 : repas

14h30 : ***Analyse des données d'enquête, data-mining et text-mining,***
par **Ludovic LEBART,**
directeur de recherches au CNRS et à l'École Nationale Supérieure des Télécommunications, ancien président de la SFdS
15h : Débats

15h30 : Pause café

16h : ***La modélisation spatiale du trafic téléphonique et modélisations,***
par **Jean BARBE,**
expert-statisticien à la Cour des Comptes européenne,
antérieurement statisticien à France Télécom
16h30 : Débats

17h : Séance de clôture

■ Institut de recherche pour le développement ■ Ecole Nationale Supérieure de Statistique et d'Économie Appliquée
■ Observatoire Économique et Statistique d'Afrique Subsaharienne ■ Société Française de Statistique
■ Association Internationale des Statisticiens d'Enquêtes

Société Française
de Statistique

AIR FRANCE

■ Société Ivoirienne de Raffinage ■ Union Européenne ■ Coopération Française ■ Banque Mondiale ■ Air France