REVUE DE STATISTIQUE APPLIQUÉE

CHRISTIAN DERQUENNE
CLÉMENCE HALLAIS

Une méthode alternative à l'approche PLS : comparaison et application aux modèles conceptuels marketing

Revue de statistique appliquée, tome 52, n° 3 (2004), p. 37-72
<http://www.numdam.org/item?id=RSA_2004__52_3_37_0>

Rev. Statistique Appliquée, 2004, **LII** (3), 37-72

UNE MÉTHODE ALTERNATIVE À L'APPROCHE PLS : COMPARAISON ET APPLICATION AUX MODÈLES CONCEPTUELS MARKETING

Christian DERQUENNE & Clémence HALLAIS

EDF R& D, 1, avenue du Général de Gaulle 92141 Clamart Cedex - France
Email : christian.derquenne@edf.fr

RÉSUMÉ

Les modèles à relations structurelles sont très utilisés dans le domaine du marketing pour construire des indicateurs de satisfaction et comprendre les leviers de la satisfaction. Ils reposent sur des modèles déterminés *a priori* par les experts du domaine et établissent des liens entre différentes facettes de la satisfaction des consommateurs. Les deux approches statistiques pour construire de tels modèles sont la méthode LISREL (LInear Structural RELationships) et l'approche PLS (Partial Least Squares). Ce papier introduit une approche alternative : RFPC (Regression on First Principal Components) et la compare aux deux autres sur des résultats d'une étude, et à l'aide d'une approche géométrique. Enfin, nous avons mis en œuvre un modèle « libre » complètement adapté aux données.

Mots-clés : *Modèles linéaires structurels, Variables latentes, LISREL, Approche PLS, Composantes principales, Marketing, Modèle ECSI (European Customer Satisfaction Index).*

ABSTRACT

The Structural Linear Models are very used in frame of marketing to make satisfaction indexes and to understand the leverage of satisfaction. These models are based on predifined *a priori* models by experts of a studied domain and allow to establish links between different aspects of customers' satisfaction. To make these models, there are two main statistical approaches : the LISREL method (LInear Structural RELationships) and PLS approach (Partial Least Squares). This paper introduces an alternative approach : RFPC (Regression on First Principal Components) and compares its to other ones on results of study and with a geometrical approach. Lastly, a free model completely suited to data is done.

Keywords : *Structural Linear Models, Latent variables, LISREL, PLS Approach, Principal components, Marketing, ECSI model (European Customer Satisfaction Index).*

1. Contexte - Objectifs

Dans le cadre de la « nouvelle économie » au niveau européen, l'évolution des comportements des différents acteurs est de plus en plus rapide. Par conséquent, des instruments de mesure de cette évolution commencent à voir le jour. Electricité de France, avec l'ouverture du marché européen de l'énergie à la concurrence, en est depuis longtemps consciente. D'ailleurs, sa Direction Marketing effectue depuis de nombreuses années des enquêtes de satisfaction auprès des différents secteurs clientèle (particuliers, professionnels, PME-PMI et grands clients), afin de détecter les points faibles des produits et services offerts par EDF. EDF met alors en place des stratégies visant à améliorer la qualité et à réaliser de nouvelles offres, qui, par conséquent, doivent permettre de mieux satisfaire et de mieux fidéliser sa clientèle. En 2000, EDF R&D et la Direction Marketing ont travaillé sur la construction d'indicateurs de satisfaction.

La construction de ce type d'indicateurs repose sur l'étude de relations structurelles définies par des relations « causales » entre des variables non observables, dites variables latentes et des variables observées, dites manifestes. Les variables latentes correspondent aux facettes ou aux composantes de la satisfaction des consommateurs; les variables manifestes ont trait aux réponses des clients à des questions concernant leur satisfaction. L'étude des liens entre les variables s'appuie sur différents modèles conceptuels marketing provenant de la théorie du processus de décision des consommateurs. Le choix du modèle est primordial, car il conditionne par la suite toute l'étude des liaisons. Par conséquent, on conçoit bien que le schéma structurel de ce type de modèle n'est pas choisi au hasard; il est fixé *a priori* par les experts du domaine. Plusieurs modèles conceptuels marketing permettant de mettre en œuvre des indices économiques ont donc été développés. Dans ce papier, nous discuterons plus précisément du modèle ECSI (European Customer Satisfaction Index) [ASL 00] qui est dérivé du modèle américain ACSI de satisfaction client.

Deux approches statistiques sont généralement utilisées pour estimer un tel modèle : la méthode LISREL (LInear Structural RELationships) [JOR 82] utilise un système d'équations structurelles estimées à l'aide du maximum de vraisemblance. Elle fournit des tests statistiques sur les relations « causales » et sur la qualité du modèle. Mais il peut exister des situations où le modèle n'est pas estimable (problème de convergence de l'algorithme). L'approche PLS (Partial Least Squares) [LOH 89, TEN 99, WOL 82] utilise un système de calculs alternés (processus itératif) entre l'estimation du modèle externe reliant les variables latentes et manifestes (composantes PLS), et celle du modèle interne rattachant les variables latentes entre elles à l'aide de régressions multiples. Cette approche peut être utilisée avec peu d'observations et beaucoup de variables. Les données peuvent être de nature différente (continues, nominales ou booléennes). En général, il n'y a pas de problème pour estimer le modèle avec cette approche. Cependant, la convergence du processus d'estimation n'a pas été démontrée théoriquement pour plus de deux variables latentes. Ces deux méthodes ont subi des évolutions différentes : LISREL est disponible dans des logiciels standards, alors que les développements informatiques concernant l'approche PLS ont été ralentis avec les décès de H. Wold et de J.B. Lohmöller [LOH 87] à la fin des années 80. Cependant, W. Chin a continué à travailler sur cette approche [CHI 98 & CHI 99] et a développé un logiciel, nommé PLS-Graph [CHI 01].

Nous proposons tout d'abord, dans ce papier, une autre approche dite « alternative » ou RFPC (Regression on First Principal Components) développée par Ch. Derquenne [DER 00]. Celle-ci repose sur le même principe que l'approche PLS. Mais les variables latentes sont estimées par les premières composantes principales issues de l'ACP (Analyse en Composantes Principales), alors que le modèle interne est également estimé par des régressions multiples. Bien que la régression après ACP soit une technique classique, l'originalité de cette nouvelle approche tient plus dans l'emploi de la régression (par l'intermédiaire de la première composante principale), que dans l'algorithme d'estimation des paramètres du modèle structurel. Il n'y a donc pas de système de calculs alternés comme dans l'approche PLS, les variables latentes sont estimées en une seule fois. Puis, nous comparons cette méthode à LISREL et à l'approche PLS sur des résultats d'une étude EDF, et à l'aide d'une approche géométrique. Enfin, l'estimation du modèle conceptuel marketing par la méthode LISREL et l'approche PLS est très fortement dépendante de la structure de ce modèle, alors que ce n'est pas le cas pour l'approche RFPC. Nous sommes alors allés encore plus loin dans la démarche. En effet, nous avons construit un modèle (dit « modèle libre ») complètement adapté aux données, dans lequel aucune hypothèse n'est faite sur le modèle externe et sur le modèle interne.

2. Un indice de satisfaction : ECSI

Nous distinguons volontairement l'indice de satisfaction ECSI de la méthode statistique permettant de l'estimer (LISREL, approche PLS et approche RFPC). En effet, un modèle marketing fait appel à l'expérience reconnue des différents concepteurs et est adapté à la demande de l'utilisateur. Ce modèle aurait très bien pu prendre une autre forme dans un domaine différent de celui de l'énergie, voire de l'électricité. Alors que les méthodes statistiques employées sont seulement des outils parmi d'autres pour estimer ce modèle conceptuel.

L'Indice de Satisfaction des Consommateurs Européens (European Consumer Satisfaction Index : ECSI) est un indicateur économique qui mesure la satisfaction des clients. Il est une adaptation de l'Indice de Satisfaction des Consommateurs Américains : ACSI. Dans ce modèle, sept variables latentes en interrelations sont introduites. Il est fondé sur des théories éprouvées et des approches sur le comportement des consommateurs. Il peut être appliqué à divers types d'industries.

Le modèle ECSI (*cf.* figure 1, ci-dessous) contient :

- Le corps du modèle, c'est-à-dire les variables latentes classiques : qualité perçue, attentes, valeur perçue, indices de satisfaction et de fidélité écrites en gras et leurs impacts mutuels, représentés par des flèches en traits pleins.
- Deux variables latentes optionnelles : image et réclamations écrites en italique et leurs impacts, sous forme de flèches en traits interrompus.

Les variables sur le coté gauche peuvent être vues comme des variables candidates à l'explication de l'indice de satisfaction et la partie droite comme un indice de performance (fidélité/réclamations). Un ensemble de variables manifestes (observables ou mesurées) est associé à chaque variable latente (par exemple, les réponses des clients à l'enquête).

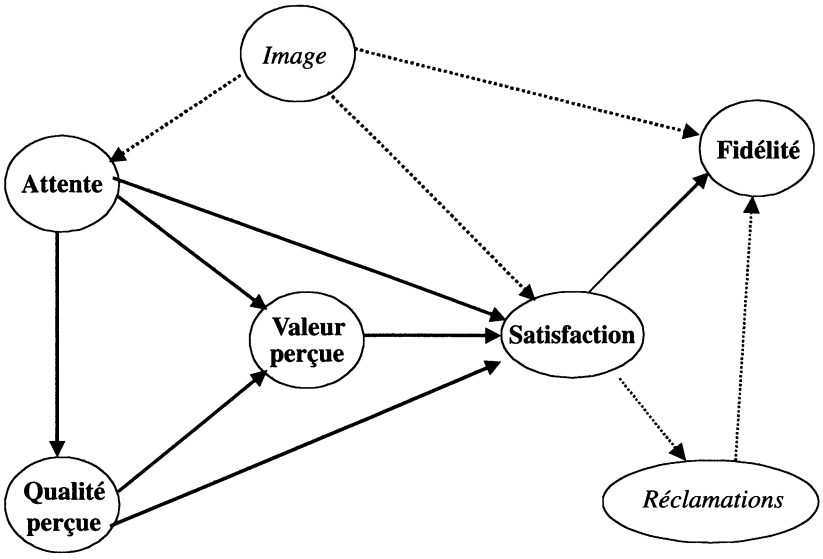

FIGURE 1
Modèle ECSI

Ce modèle a été utilisé sur les données EDF dans lesquelles, il n'y a pas eu d'informations relatives aux réclamations dans ce questionnaire, on considérera donc ici le modèle ECSI sans réclamations. Enfin, signalons un point très important : ce modèle est construit sur des bases d'un a priori important de forme et de liens entre les variables latentes.

3. Trois méthodes pour estimer des relations structurelles : LISREL, PLS et RFPC

Dans cette partie, nous donnons le principe des relations structurelles. Puis, nous rappelons assez brièvement les méthodes LISREL et PLS. Enfin, nous introduisons la méthode alternative (RFPC) proposée dans ce papier. Pour chaque méthode, nous fournissons les résultats sur le modèle ECSI.

3.1. Les relations structurelles

Un système d'équations structurelles est composé de deux types de variables :

– les variables manifestes

– les variables latentes.

Les variables manifestes (ou points de mesure) correspondent à des facteurs observables (par exemple, la taille, le PNB par habitant, le lieu d'habitation....etc..), elles sont de nature qualitative ou quantitative. Dans notre cas, ce sont les notes attribuées à chaque affirmation constituant l'enquête (par exemple r1, r2, sat1 ...).

Chaque variable manifeste est l'expression observable d'une variable latente de nature subjective (par exemple la qualité de vie, l'instabilité politique...) qui ne peut pas être quantifiée. Plusieurs variables manifestes sont en général utilisées pour décrire une même variable latente. On définit les relations causales entre variables manifestes et latentes de la façon suivante :

$$\forall h = 1, \ldots p_i, \qquad X_{ih} = \gamma_{ih} + d_{ih} L_i \tag{1}$$

où les X_{ih} sont les variables manifestes, au nombre de p_i, et L_i la variable latente dont elles sont dépendantes et où γ_{ih} est supposée de moyenne nulle et non corrélée à la variable latente L_i.

Puis viennent les relations entre les différentes variables latentes qui peuvent s'écrire :

$$L_j = \sum_{i \neq j} c_{ij} L_i + K_j \tag{2}$$

où K_j est une erreur supposée de moyenne nulle et non corrélée aux variables latentes L_i. Si la variable L_i n'influe pas sur la variable L_j le coefficient c_{ij} est alors nul. On représente un schéma d'équations structurelles de la façon suivante (*cf.* figure 2) :

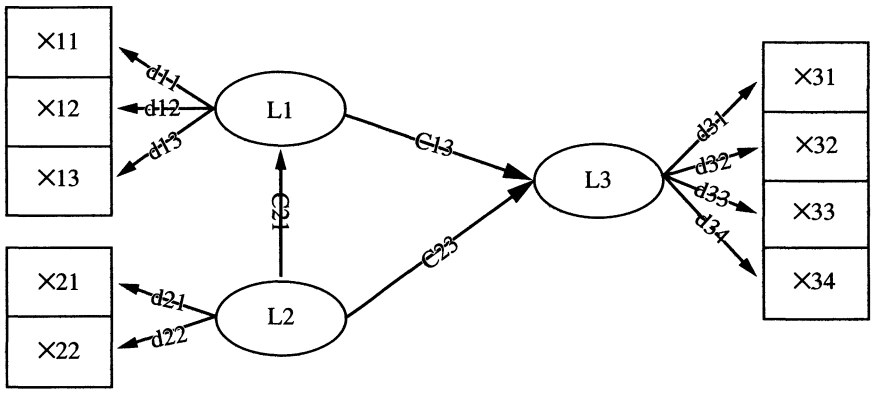

FIGURE 2
Schéma de relations structurelles

Les carrés représentent ici les variables manifestes et les ovales : les variables latentes.

Le système d'équations structurelles s'écrit, pour le schéma précédent, comme suit :

Équations entre les variables manifestes et les variables latentes :

$$\begin{array}{lll} X_{11} = \gamma_{11} + d_{11} L_1 & X_{21} = \gamma_{21} + d_{21} L_2 & X_{31} = \gamma_{31} + d_{31} L_3 \\ X_{12} = \gamma_{12} + d_{12} L_1 & X_{22} = \gamma_{22} + d_{22} L_2 & X_{32} = \gamma_{32} + d_{32} L_3 \\ X_{13} = \gamma_{13} + d_{13} L_1 & & X_{33} = \gamma_{33} + d_{33} L_3 \\ & & X_{34} = \gamma_{34} + d_{34} L_3 \end{array} \tag{3}$$

Équations entre variables latentes :

$$\begin{aligned} L_1 &= c_{21}L_2 + K_1 \\ L_3 &= c_{13}L_1 + c_{23}L_2 + K_3 \end{aligned} \tag{4}$$

On distingue également des variables latentes exogènes (explicatives et non expliquées) (par exemple L_2) et des variables latentes endogènes (expliquées) (par exemple L_1 et L_3). Généralement une des variables latentes endogènes est dite variable latente cible et c'est sur elle que vont converger les autres flèches (c'est le cas de la variable L_3 de notre schéma), elle est dépendante des autres variables latentes et est donc non explicative, c'est le cas dans notre questionnaire de la variable représentant la fidélité.

Il s'agit également de voir quel est le lien entre toutes les autres variables et plus précisément d'essayer d'estimer les coefficients d_{ij} et c_{ij}. Puis on regardera pour chaque variable latente quelles sont les variables manifestes qui lui sont le plus corrélées donc qui ont un impact plus important dans sa détermination.

Voyons maintenant quelle est la forme du modèle ECSI, sous forme de relations structurelles, à l'aide des équations (2) fournies précédemment.

Modèle ECSI :

Attente $= \beta_{a0} + \beta_{a1}$Image $+\zeta_a$

Qualité Perçue $= \beta_{q0} + \beta_{q1}$Attente $+\zeta_q$

Valeur Perçue $= \beta_{v0} + \beta_{v1}$Attente $+\beta_{v2}$Qualité Perçue $+\zeta_v$

Satisfaction $= \beta_{s0} + \beta_{s1}$Attente $+\beta_{s2}$Qualité Perçue $+\beta_{s3}$Valeur Perçue $+\beta_{s4}$Image $+\zeta_s$

Fidélité $= \beta_{f0} + \beta_{f1}$Satisfaction $+\beta_{f2}$Image $+\zeta_f$

Dans ce modèle, seule la variable : ***image*** est exogène, alors que les variables : ***qualité perçue***, ***attentes***, ***valeur perçue***, ***satisfaction et fidélité*** sont endogènes.

3.2. La méthode LISREL

Cette méthode développée par K.G. Jöreskog [JOR 79] permet l'étude d'équations structurelles dans un modèle avec variables latentes.

Une équation structurelle peut se représenter sous forme matricielle de la façon suivante :

$$\eta = B\eta + \Gamma\xi + \zeta \tag{5}$$

où η : vecteur $(m \times 1)$ des variables latentes endogènes

ξ : vecteur $(n \times 1)$ des variables latentes exogènes

ζ : vecteur $(m \times 1)$ des erreurs

B : matrice $(m \times m)$ des coefficients pour les variables latentes endogènes

Γ : matrice $(m \times n)$ des coefficients pour les variables latentes exogènes

Φ : matrice de covariance de ξ

Ψ : matrice de covariance de ζ

et ξ et ζ sont non corrélées.

Les variables manifestes sont reliées aux variables latentes dans le modèle de mesure suivant :

$$\begin{aligned} y &= \Lambda_y \eta + \varepsilon \\ x &= \Lambda_x \xi + \delta \end{aligned} \tag{6}$$

où,

y : vecteur $(p \times 1)$ des variables manifestes relatives aux variables latentes endogènes centrées

x : vecteur $(q \times 1)$ des variables manifestes relatives aux variables latentes exogènes centrées

Λ_x : matrice $(q \times n)$ des coefficients reliant x à ξ

Λ_y : matrice $(p \times m)$ des coefficients reliant y à η

ε : vecteur $(p \times 1)$ des erreurs

δ : vecteur $(q \times 1)$ des erreurs

Θ_ε : matrice de covariance de ε

Θ_δ : matrice de covariance de δ

et ε et η sont non corrélées,

ξ et δ sont non corrélées,

ε , δ et ζ sont non corrélées,

η et δ sont non corrélées,

ε et ξ sont non corrélées.

La procédure d'estimation provient des relations entre la matrice de covariance des variables observées et les paramètres des équations structurelles.

Soit θ le vecteur des paramètres à estimer (θ contient t paramètres : les coefficients des matrices B, Γ, Φ, Ψ, Λ_y, Λ_x, Θ_ε et Θ_δ).

L'hypothèse de base du modèle général d'équations structurelles est :

$$\Sigma = \Sigma(\theta) \tag{7}$$

où Σ est la matrice de covariance des x et y, et $\Sigma(\theta)$ la matrice de covariance écrite comme fonction des paramètres du modèle. On suppose également que $(I - B)^{-1}$ est inversible. On décompose $\Sigma(\theta)$ en trois parties : la matrice de covariance de y : $\Sigma_{yy}(\theta)$, la matrice de covariance de x : $\Sigma_{xx}(\theta)$ et la matrice de covariance de x avec y : $\Sigma_{xy}(\theta)$.

On a :

$$\begin{aligned}\Sigma_{yy}(\theta) &= E(yy') \\ &= E[(\Lambda_y\eta + \varepsilon)(\eta'\Lambda_y' + \varepsilon')] \\ &= \Lambda_y E(\eta\eta')\Lambda_y' + \Theta_\varepsilon\end{aligned}$$

puisque ε et η sont non corrélées, or d'après la formule (5), $\eta = (I - B)^{-1}(\Gamma\xi + \zeta)$ donc en remplaçant par cette valeur on obtient :

$$\Sigma_{yy}(\theta) = \Lambda_y(I - B)^{-1}(\Gamma\Phi\Gamma' + \Psi)[(I - B)^{-1}]'\Lambda_y' + \Theta_\varepsilon,$$

puisque ξ et ζ sont non corrélées.

$$\begin{aligned}\Sigma_{yx}(\theta) &= E(yx') \\ &= E[(\Lambda_y\eta + \varepsilon)(\xi'\Lambda_x' + \delta')] \\ &= \Lambda_y E(\eta\xi')\Lambda_x' \\ &= \Lambda_y(I - B)^{-1}\Gamma\Phi\Lambda_x' \text{ en utilisant la forme réduite de } \eta\end{aligned}$$

puisque ε et δ sont non corrélées, de même pour ε et ξ, et η et δ.

Enfin $\Sigma_{xx}(\theta) = E(xx') = \Lambda_x\Phi\Lambda_x' + \Theta_\delta$, puisque que ξ et δ sont non corrélées.

D'où

$$\begin{aligned}\Sigma(\theta) &= \begin{bmatrix}\Sigma_{yy}(\theta) & \Sigma_{yx}(\theta) \\ \Sigma_{xy}(\theta) & \Sigma_{xx}(\theta)\end{bmatrix} \\ &= \begin{bmatrix}\Lambda_y(I - B)^{-1}(\Gamma\Phi\Gamma' + \Psi)[(I - B)^{-1}]'\Lambda_y' + \Theta_\varepsilon & \Lambda_y(I - B)^{-1}\Gamma\Phi\Lambda_x' \\ \Lambda_x\Phi\Gamma'[(I - B)^{-1}]'\Lambda_y' & \Lambda_x\Phi\Lambda_x'\Theta_\delta\end{bmatrix}\end{aligned}$$

De plus, on peut estimer la matrice de covariance des variables manifestes $\Sigma(\delta)$ par la matrice de covariance empirique que l'on notera S. En identifiant les matrices $\Sigma(\theta)$ et S terme à terme, on obtient un système de $\frac{1}{2}(p + q)(p + q + 1)$ équations à t inconnues.

Il faut alors trouver, quand cela est possible *i.e.* quand le modèle est identifiable, $\hat{\Sigma}(\theta)$ (estimateur de $\Sigma(\theta)$) telles que $\hat{\Sigma}(\theta)$ et S soient le plus proche possible. Il faut donc définir une notion de « proximité » entre ces deux matrices : pour cela on cherche une fonction $F(S, \Sigma(\theta))$ et il faut trouver $\hat{\Sigma}(\theta)$ telle que $F(S, \Sigma(\theta))$ soit minimum en $\hat{\Sigma}(\theta)$.

Plusieurs fonctions peuvent convenir, elles doivent cependant vérifier les conditions :

$F\big(S, \Sigma(\theta)\big)$ est un scalaire

$F\big(S, \Sigma(\theta)\big) \geqslant 0$

$F\big(S, \Sigma(\theta)\big) = 0 \Leftrightarrow S = \Sigma(\theta)$

$F\big(S, \Sigma(\theta)\big)$ est continue en S et en $\Sigma(\theta)$

Selon Browne, minimiser de telles fonctions d'ajustement conduit à des estimateurs consistants de θ.

Une des fonctions qui convient est la fonction définie par :

$$F\big(S, \Sigma(\theta)\big) = \log|\Sigma(\theta)| + tr\big(S \cdot \Sigma^{-1}(\theta)\big) - \log|S| - (p+q) \qquad (8)$$

où l'on suppose généralement que S et $\Sigma(\theta)$ sont définies positives, fonction qui n'est autre que la log-vraisemblance, si on se place dans un cadre gaussien. De plus, $|A|$ désigne le déterminant de la matrice carrée A. On vérifie facilement que cette fonction s'annule en $\Sigma(\theta) = S$ et on recherche, s'il existe, le minimum de la fonction en calculant ses dérivées partielles par rapport à chacun des paramètres inconnus. Bien souvent des solutions explicites ne sont pas trouvées; il faut alors faire appel à une procédure numérique itérative. De telles procédures sont développées dans [BOL 89].

Modèle ECSI [1] :

Seuls les résultats issus de la méthode LISREL sur les coefficients de détermination R^2 sont présentés ici. Les coefficients et les niveaux de signification entre les variables latentes ne sont pas donnés car ils sont confidentiels.

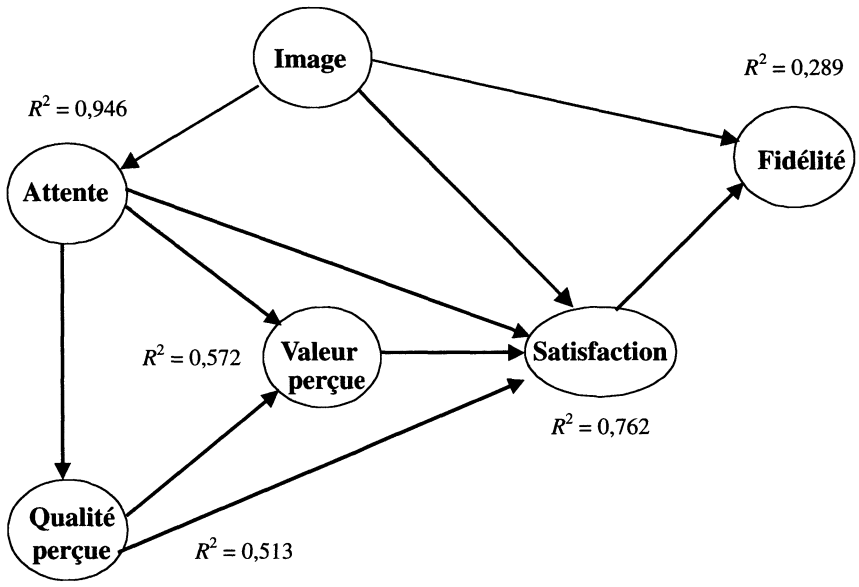

FIGURE 3
Modèle ECSI avec la méthode LISREL

[1] La méthode LISREL a été appliquée à l'aide de la procédure CALIS du Système SAS [SAS 01].

3.3. L'approche PLS

Dans le cadre de la modélisation de relations structurelles entre variables latentes, l'approche PLS est une méthode d'analyse des données proposée par H. Wold [WOL 82] et qui permet d'étudier J blocs de variables observées sur les mêmes individus. Cette méthode est décrite en détail dans [TEN 99]. Les développements informatiques concernant l'approche PLS ont été ralentis avec les décès de Wold et Lohmöller qui ont développé le logiciel LVPLS 1.8 disponible sous DOS [LOH 87]. Cependant, W. Chin a continué a travaillé sur cette approche [CHI 98 & CHI 99] et a développé un logiciel, nommé PLS-Graph [CHI 01].

Il y a J groupes de variables : $X_j = (x_{j1}, \ldots, x_{jpj})$, observées sur n individus. Les variables x_{jh} sont les variables manifestes et correspondent aux notes données à chaque question de l'enquête. Chaque groupe de variables constitue l'expression observable d'une variable latente ξ_j centrée réduite.

Les variables manifestes sont reliées à leur variable latente ξ_i par l'équation linéaire :

$$x_{jh} = \pi_{jh}\xi_j + \varepsilon_{jh} \tag{9}$$

où ε_{jh} est une erreur supposée de moyenne nulle et non corrélée à la variable latente ξ_j.

Le phénomène étudié est décrit par des relations structurelles entre les variables latentes de la forme :

$$\xi_j = \sum_{i \neq j} \beta_{ij}\xi_i + \eta_j \tag{10}$$

où η_j est supposée de moyenne nulle et non corrélée aux variables latentes ξ_i. On doit nécessairement préciser le sens de la corrélation entre la variable latente dépendante ξ_j et les autres variables latentes ξ_i apparaissant dans cette même équation :

$$\begin{aligned} c_{ij} = c_{ji} &= +1 \text{ si la corrélation entre } \xi_i \text{ et } \xi_j \text{ est positive} \\ &= -1 \text{ si elle est négative.} \end{aligned}$$

où les c_{ij} interviennent plus loin dans l'équation (12).

Si les variables ξ_i et ξ_j ne sont pas supposées liées entre elles, on pose $c_{ij} = c_{ji} = 0$: les coefficients β_{ij} et β_{ji} sont, dans ce cas, structurellement nuls.

Les variables latentes ξ_j sont estimées par des combinaisons linéaires Y_j des variables x_{jh} :

$$Y_j = \sum_h w_{jh}x_{jh} = X_j w_j \tag{11}$$

avec w_j le vecteur colonne des poids w_{jh}. On impose aussi à Y_j d'être centrée et réduite.

On définit également une autre approximation Z_j de ξ_j construite à l'aide des estimations Y_i des variables latentes ξ_i liées à ξ_j :

$$Z_j \propto \sum_{i \neq j} c_{ij} Y_i \tag{12}$$

où le signe $\propto$ signifie que la variable située à gauche de ce signe est obtenue par réduction de la variable située à droite.

Les variables Y_j et Z_j peuvent être considérées comme des approximations respectivement externes et internes de ξ_j. En reliant ces deux approximations, Wold obtient des conditions de stationnarité qui permettent de déterminer les variables Y_j. Les équations de stationnarité sont résolues par un processus itératif dont la convergence est prouvée dans le cas de deux groupes et constatée dans la pratique pour des situations plus générales.

Un des modes de relation (mode A, dans [TEN 99]) entre les deux approximations Y_j et Z_j de ξ_j est le suivant :

$$Y_j \propto \sum_h \mathrm{cov}(x_{jh}, Z_j) x_{jh} \tag{13}$$

qui peut s'écrire matriciellement :

$$Y_j \propto X_j X_j' Z_j \tag{14}$$

La condition de stationnarité pour une variable Y_j dans ce mode est :

$$Y_j \propto X_j X_j' \sum_{i \neq j} c_{ij} Y_i \tag{15}$$

L'algorithme itératif proposé par Wold est le suivant :

- à l'étape initiale, on part de variables Y_i arbitrairement fixées
- on obtient à l'aide de l'équation (15) de nouvelles valeurs de ces variables
- on itère jusqu'à convergence et on obtient les estimations Y_j des variables latentes ξ_j

Les vecteurs-poids w_j sont déduits de (15).

Les paramètres du modèle définis par les équations (9) et (10) sont estimées par régression en remplaçant les variables latentes ξ_j par leur estimation Y_j.

Les équations (9) sont estimées par régression simple de chaque x_{jh} sur Y_j :

$$x_{jh} \approx p_{jh} Y_j \tag{16}$$

Les équations (10) sont estimées par régression multiple de Y_j sur les variables Y_i qui lui sont structurellement liées (*i.e.* $c_{ij} \neq 0$) :

$$Y_j \approx \sum_{i \neq j} b_{ij} Y_i \tag{17}$$

L'approche PLS est donc assez simple à mettre en œuvre même si elle impose au départ de fixer des hypothèses sur les liens entre les variables et surtout leur sens ce qui n'est pas forcément évident avant l'analyse.

L'utilisation du programme LVPLS 1.8 écrit par Wold et Lohmöller [LOH 87] nous a permis de bien nous approprier les aspects pratiques de cette méthode. En effet, nous l'avons appliqué sur différents jeux de données, sur lesquels, nous avons essayé différentes options offertes pas ce programme (choix du mode de relation, métrique des données,...). On obtient en sortie, en plus des coefficients de régression, la matrice des poids de chaque question sur la variable latente à laquelle elle se rapporte, on peut donc déterminer l'ordre d'importance de chaque variable manifeste sur une même variable latente.

Modèle ECSI :

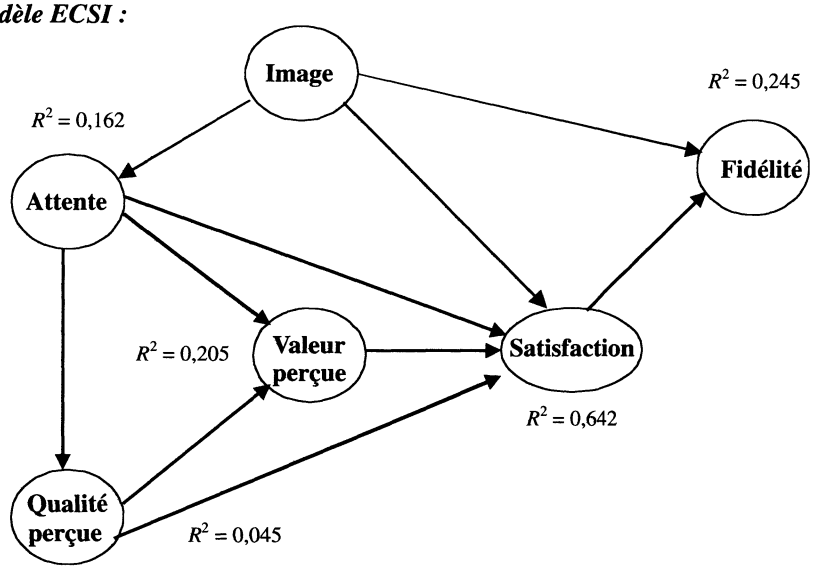

FIGURE 4
Modèle ECSI avec l'approche PLS

3.4. L'approche alternative : Regression on First Principal Components

Une approche alternative : Regression on First Principal Components (RFPC) aux deux précédentes a été développée par Ch. Derquenne en 2000 [DER 00]. Comme on va le constater, cette approche est assez similaire à l'approche PLS, mais utilise des

outils statistiques plus simples, et a l'avantage d'être plus souple et plus contrôlable du point de vue du modèle conceptuel fixé *a priori*.

On dispose toujours de J blocs de variables : $X_j = (x_{j1}, \ldots, x_{jpj})$ observés sur n individus, où les x_{jh} représentent aussi les variables manifestes. Et chaque groupe de variables est également l'expression observable d'une variable latente ξ_j centrée et réduite.

Comme dans l'approche PLS, l'approche proposée contient deux étapes correspondant à l'estimation externe et à l'estimation interne des variables latentes. Cependant, contrairement à l'approche PLS, chaque variable latente est estimée une seule fois dans l'étape d'estimation externe. En effet, il n'y a pas d'algorithme itératif estimant chaque variable latente jusqu'à convergence. Cette variable est estimée à l'aide de la première composante principale réduite Y_{1j}^* issue de l'ACP (Analyse en Composantes Principales) de chaque paquet de variables associé à ξ_j. En effet, la première composante principale résume le maximum de variance expliquée. Par conséquent, il est tout à fait naturel d'avoir choisi cette approche. Cette variable latente a la forme suivante :

$$Y_{1j} = \sum_{h=1}^{pj} \alpha_{1jh} x_{jh} = X_j \alpha_{1j} \tag{18}$$

Le vecteur colonne α_{1j} représente les poids respectifs de chaque variable dans le sous-ensemble j de variables. Ces poids correspondent au vecteur unitaire v_{1j} divisé par la racine carrée de la première valeur propre λ_{1j}, où v_{1j} correspond au vecteur propre normé de la matrice des corrélations du bloc j de variables, associé à la première valeur propre λ_{1j}. On sait par ailleurs que la corrélation $\mathrm{cor}(x_{jh}, Y_{1j}^*)$ entre la première composante principale et chaque variable x_{jh} en ACP est égale à $\sqrt{\lambda_{1j}} v_{1jh}$, où v_{1jh} est la $h^{\text{ième}}$ composante de v_{1j}. Par conséquent, les poids α_{1jh} sont égaux à $v_{1jh}/\sqrt{\lambda_{1j}} = \mathrm{cor}(x_{jh}, Y_{1j}^*)/\lambda_{1j} = \mathrm{cov}(x_{jh}, Y_{ij}^*)/(\lambda_{1j} \times \sigma(x_{jh}))$, où $\sigma(x_{jh})$ est l'écart-type de x_{jh}, alors que Si x_{jh} est centrée-réduite (notée x_{jh}^*), alors les poids respectifs pour RFPC et PLS sont les suivants :

$$\alpha_{1jh} = \mathrm{cor}(x_{jh}, Y_{1j}^*)/\lambda_{1j} \quad \text{et} \quad w_{jh}^* = \mathrm{cov}(x_{jh}, Z_j) \tag{19}$$

L'estimation interne du modèle structurel correspond à l'approximation Z_j de ξ_j qui est construite à l'aide des estimations Y_{1i} des variables latentes ξ_i liées à ξ_j :

$$Z_j \propto \sum_{i \neq j} c_{ij} Y_{1j} \tag{20}$$

où le signe $\propto$ signifie que la variable située à gauche de ce signe est obtenue par réduction de la variable située à droite. Les équations (20) sont estimées par régression multiple de Y_{1j} sur les variables Y_{1i} qui lui sont structurellement liées, comme dans le schéma structurel proposé par Lohmöller [LOH 89] :

$$Y_j = \sum_{i \neq j} b_{ij} Y_{1i} \tag{21}$$

L'estimation interne est également réalisée en une seule fois. On peut alors réaliser des tests de nullité des coefficients dans chaque équation de régression, ce qui permet de remettre en cause des relations structurelles imposées entre variables latentes. On peut également calculer les corrélations de chaque question avec sa composante principale, ce qui nous permet d'obtenir l'ordre d'impact des différentes variables manifestes qui composent une même variable latente et de le comparer avec les résultats fournis par l'approche PLS.

Contrairement aux deux autres méthodes, celle-ci ne présuppose pas un modèle trop rigide; en effet une fois fixés les différents groupes et les questions qui les composent, l'estimation des variables latentes au niveau de chaque individu est également déterminée et on est alors libre de faire les régressions qui nous intéressent. Ce qui n'est pas le cas dans l'approche PLS où l'estimation des variables latentes se fait précisément en fonction du modèle. Cependant pour comparer la méthode RFPC avec ces autres méthodes, il faut bien sûr se fixer un modèle et effectuer les régressions correspondantes.

Modèle ECSI :

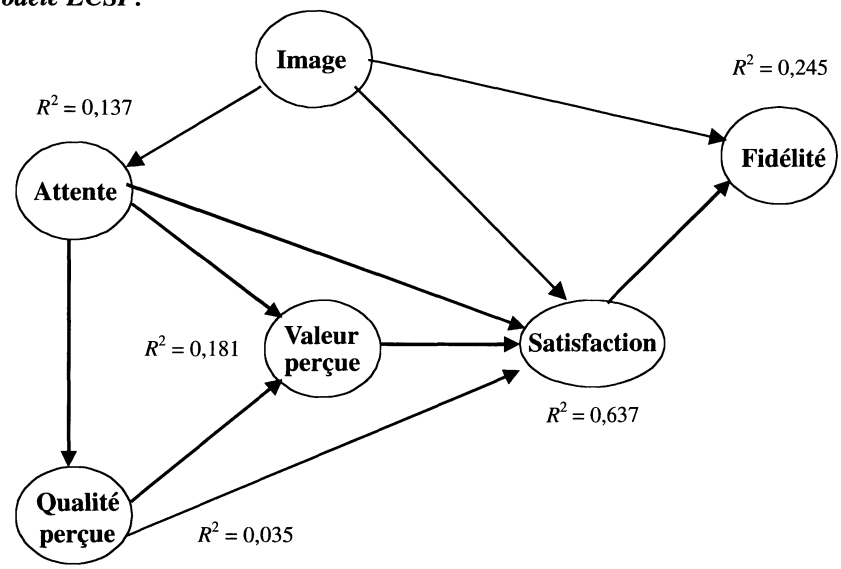

FIGURE 5
Modèle ECSI avec l'approche alternative

Les résultats obtenus avec l'approche alternative sont tout à fait similaires à ceux de l'approche PLS, en terme d'impact entre variables latentes. On remarquera également que les R^2 sont très proches de ceux de PLS.

4. Comparaison des méthodes

Dans cette partie, nous confrontons tout d'abord les résultats obtenus par les trois méthodes, puis nous comparons l'approche PLS et l'approche alternative à l'aide d'une approche géométrique. Sachant qu'il existe de nombreux résultats théoriques et empiriques de comparaison entre la méthode LISREL et l'approche PLS [JOR 82, TEN 01], nous n'en discuterons pas dans ce papier.

4.1. Comparaison des résultats

Nous comparons tout d'abord les coefficients standardisés des relations entre les variables manifestes et leurs variables latentes (estimation externe), et les corrélations associées, puis nous examinons les intervalles de confiance des paramètres liés aux relations entre variables latentes (estimation interne), ainsi que les coefficients de détermination de chacune des variables latentes. Enfin, nous calculons les corrélations respectives pour une même variable latente entre les trois méthodes.

4.1.1. Estimation externe

Nous reprenons dans cette comparaison l'approche de C. Fornell [FOR 92] consistant à résumer chaque bloc lié à une variable latente à l'aide d'une moyenne pondérée de ses variables manifestes. Cette approche a été utilisée dans [TEN 01] sur la téléphonie mobile pour comparer l'approche PLS et la méthode LISREL et a fourni des résultats très satisfaisants.

Les moyennes pondérées sont les suivantes :

Approche LISREL :

$$\widetilde{\xi}_j = \sum_h \lambda_{jh} x_{jh} \Big/ \sum_h \lambda_{jh} \tag{22}$$

où les λ_{jh} correspondent au vecteur Λ_x pour une variable latente exogène, et Λ_y pour une variable latente endogène (*cf.* formules (6)).

Approche PLS :

$$\widetilde{\xi}_j = \sum_h \mathrm{cov}(x_{jh}, \xi_j) X_{jh} \Big/ \sum_h \mathrm{cov}(x_{jh}, \xi_j) \tag{23}$$

Approche RFPC :

$$\widetilde{\xi}_j = \sum_h \alpha_{1jh} x_{jh} \Big/ \sum_h \alpha_{1jh} \tag{24}$$

où $\alpha_{1jh} = v_{1jh}/\sqrt{\lambda_{1j}}$ et v_{1jh} est la $h^{\text{ième}}$ composante du premier vecteur axial factoriel dans l'ACP normée du $j^{\text{ième}}$ groupe de variables manifestes, composante

associée à la valeur propre λ_{1j} et à la variable manifeste X_{jh} de la variable latente ξ_j. Le calcul de ces trois formules suppose que la pondération totale associée à chaque méthode soit non nulle.

La comparaison des poids respectifs des trois méthodes est réalisée en construisant des poids relatifs égaux :

Approche LISREL :

$$\lambda_{jh} \Big/ \sum_h \lambda_{jh} \tag{25}$$

Approche PLS :

$$\mathrm{cov}(x_{jh}, \xi_j) \Big/ \sum_h \mathrm{cov}(x_{jh}, \xi_j) \tag{26}$$

Approche RFPC :

$$\alpha_{jh} \Big/ \sum_h \alpha_{jh} \tag{27}$$

On peut constater sur le tableau 1 suivant, que les poids relatifs entre les trois méthodes sont plus ou moins ressemblants selon la variable latente. Afin de résumer ce tableau, nous avons calculé pour chacune des variables latentes un indice de proximité (moyenne des valeurs absolues de la différence entre les poids relatifs) entre LISREL, PLS et RFPC (*cf.* tableau 2) Ces résultats sont bien évidemment complètement empiriques; ils sont donc seulement valables pour cette étude. Pour la qualité perçue, la valeur perçue, l'image et la satisfaction : l'approche PLS et l'approche RFPC sont les plus voisines, et LISREL et PLS sont les plus éloignées, sauf pour la satisfaction. Pour l'attente, LISREL et RFPC sont les plus proches, alors que l'approche PLS et l'approche RFPC sont les plus distantes. La raison de la plus au moins grande proximité entre PLS et l'approche RFPC sera expliquée à l'aide de l'approche géométrique dans le paragraphe 4.2.

Le tableau suivant (tableau 3) fournit les corrélations entre les variables manifestes et leur variable latente. On peut constater que les corrélations entre l'approche PLS et l'approche RFPC sont relativement ressemblantes, de même que leur ordre. Cependant, ceci ne se vérifie pas pour la variable latente concernant les attentes. Par contre, les corrélations issues de l'approche PLS (et de l'approche alternative) sont nettement plus élevées que celle de la méthode LISREL. En effet, comme le soulignent les auteurs dans [TEN 01], dans l'approche PLS, les variables latentes sont estimées sous la contrainte d'appartenir à l'espace engendré par leurs variables manifestes. Puis les paramètres du modèle sont calculés par régression simple ou multiple. Alors que dans la méthode LISREL, les paramètres du modèle sont estimés par maximum de vraisemblance et peu de contraintes sont imposées aux variables latentes. En d'autres termes, l'estimation des variables latentes ne joue aucun rôle dans l'estimation du modèle. Ceci a pour effet que les équations structurelles sont

TABLEAU 1
Poids des variables manifestes ECSI

Variables manifestes		**Approche LISREL**		**Approche PLS**		**Approche RFPC**	
		Poids	*Poids relatifs*	*Poids*	*Poids relatifs*	*Poids*	*Poids relatifs*
b1		1,000	0,181	0,372	0,234	0,257	0,163
b2		0,968	0,175	0,349	0,220	0,201	0,127
b3	**ATTENTE**	0,898	0,162	0,225	0,142	0,289	0,183
b4		0,778	0,141	0,138	0,087	0,235	0,149
b5		0,894	0,161	0,266	0,167	0,299	0,189
b6		0,999	0,180	0,238	0,150	0,299	0,189
r1		1,000	0,053	0,113	0,063	0,104	0,058
r2		1,253	0,066	0,125	0,070	0,112	0,062
r3		1,181	0,062	0,135	0,076	0,115	0,064
r4		0,769	0,041	0,082	0,046	0,074	0,041
c1		1,234	0,065	0,115	0,065	0,098	0,055
c2		1,186	0,063	0,096	0,054	0,110	0,061
c3		1,150	0,061	0,102	0,057	0,101	0,056
f1	**QUALITÉ**	0,853	0,045	0,068	0,038	0,087	0,049
f2		0,902	0,048	0,086	0,048	0,097	0,054
f3	**PERÇUE**	0,797	0,042	0,086	0,048	0,071	0,040
f4		0,809	0,043	0,099	0,056	0,102	0,057
i1		0,785	0,042	0,085	0,048	0,084	0,047
i2		0,996	0,053	0,111	0,062	0,110	0,062
i3		0,838	0,044	0,100	0,056	0,098	0,055
t1		1,111	0,059	0,080	0,045	0,093	0,052
t3		1,080	0,057	0,052	0,029	0,075	0,042
t4		1,194	0,063	0,080	0,045	0,083	0,046
t5		0,809	0,043	0,083	0,050	0,094	0,052
t6		0,972	0,051	0,077	0,043	0,081	0,045
co1	**VALEUR**	1,000	0,159	0,394	0,251	0,345	0,222
co2		1,548	0,247	0,355	0,226	0,416	0,268
pv1	**PERÇUE**	2,360	0,376	0,537	0,342	0,473	0,305
pv2		1,366	0,218	0,283	0,180	0,318	0,205
cf1		1,000	0,242	0,301	0,218	0,285	0,207
cf2		0,874	0,211	0,245	0,200	0,295	0,215
im1	**IMAGE**	0,923	0,223	0,312	0,227	0,310	0,226
im2		0,347	0,084	0,169	0,122	0,163	0,118
im3		0,995	0,240	0,321	0,233	0,321	0,234
sat1	**SATISFAC-**	1,000	0,442	0,549	0,406	0,500	0,368
sat2	**TION**	0,540	0,239	0,383	0,284	0,400	0,295
sat3		0,721	0,319	0,419	0,310	0,458	0,337
fid1	**FIDÉLITÉ**	1	1	1	1	1	1

TABLEAU 2
Moyennes des valeurs absolues des différences

Variables latentes	**LISREL *vs* PLS**	**LISREL *vs* RFPC**	**PLS *vs* RFPC**
Attente	0,035	0,022	0,055
Qual. perçue	0,009	0,007	0,005
Val. perçue	0,045	0,042	0,033
Image	0,017	0,016	0,006
Satisfaction	0,030	0,049	0,025
Fidélité	0,000	0,000	0,000

plus significatives dans LISREL que dans l'approche PLS, c'est-à-dire que les R^2 sont plus élevés, alors que les corrélations entre variables manifestes et variables latentes sont plus fortes en PLS, qu'en LISREL.

4.1.2. Estimation interne

Nous avons examiné également les valeurs des coefficients de régression entre les variables latentes, ainsi que leurs intervalles de confiance, mais nous ne pouvons pas fournir ces résultats car ils sont confidentiels. De plus, signalons qu'il est seulement possible de comparer les intervalles de l'approche PLS et l'approche alternative, car les coefficients issus de LISREL ne sont pas soumis aux mêmes contraintes. Cependant, la comparaison avec LISREL est valable pour les tests statistiques sur les liens entre variables latentes. Les résultats de comparaison montrent que tous les intervalles de confiance se chevauchent très largement. De plus, dans LISREL les liens entre valeur perçue et qualité perçue, et satisfaction et image ne sont pas significatifs, alors qu'ils le sont pour PLS et RFPC.

Le tableau 4 vérifie bien le fait que les coefficients de détermination R^2 associés à chaque variable latente sont plus élevés dans la méthode LISREL que dans l'approche PLS qui eux-mêmes (ce qui est logique) sont plus grands que ceux de l'approche alternative. Enfin, le R^2 pour l'image est nul pour les trois méthodes, car elle est la seule à ne pas être endogène.

4.1.3. Corrélations des variables latentes entre les trois méthodes

Afin de comparer les méthodes LISREL, PLS et RFPC, nous avons calculé les corrélations, pour une même variable latente [2], entre ces trois méthodes (*cf.* tableau 5). Un résultat qui était aussi apparu dans [TEN 01] est également présent dans cette étude : les corrélations entre mêmes variables latentes sont très fortes pour LISREL et PLS (toutes supérieures à 0,96). De plus, l'approche alternative a des corrélations plus élevées avec l'approche PLS sauf pour attente, ce qui est relativement logique

[2] La variable latente «Fidélité» n'apparaît pas dans les résultats car il n'y a qu'une seule variable manifeste. Par conséquent les coefficients de corrélations associés sont tous égaux à l'unité.

TABLEAU 3
Corrélations entre les variables manifestes et leurs variables latentes

Variables manifestes	**LISREL**	**PLS**	**RFPC**
b1	0,380	0,695	0,606
b2	0,342	0,608	0,474
b3 **ATTENTE**	0,306	0,635	0,681
b4	0,226	0,451	0,553
b5	0,328	0,643	0,705
b6	0,312	0,645	0,706
r1	0,598	0,624	0,606
r2	0,645	0,667	0,651
r3	0,680	0,687	0,668
r4	0,393	0,438	0,429
c1	0,547	0,585	0,573
c2	0,593	0,626	0,639
c3	0,560	0,593	0,588
f1 **QUALITÉ**	0,462	0,498	0,508
f2	0,515	0,556	0,565
f3 **PERÇUE**	0,369	0,425	0,413
f4	0,555	0,596	0,596
i1	0,442	0,495	0,491
i2	0,610	0,650	0,644
i3	0,547	0,577	0,574
t1	0,483	0,518	0,543
t3	0,360	0,405	0,434
t4	0,414	0,471	0,484
t5	0,494	0,528	0,546
t6	0,426	0,458	0,470
co1 **VALEUR**	0,368	0,603	0,558
co2	0,489	0,634	0,675
pv1 **PERÇUE**	0,679	0,802	0,767
pv2	0,294	0,448	0,515
cf1	0,656	0,731	0,723
cf2	0,652	0,738	0,749
im1 **IMAGE**	0,713	0,788	0,787
im2	0,331	0,418	0,413
im3	0,737	0,815	0,815
sat1 **SATISFAC-**	0,737	0,834	0,807
sat2 **TION**	0,358	0,635	0,645
sat3	0,558	0,715	0,739
fid1 **FIDÉLITÉ**	1	1	1

puisqu'il y a une certaine ressemblance du mode d'estimation entre elles. En effet, si la première composante principale et la composante PLS sont proches, les régressions multiples, et donc les corrélations seront plus fortes. Enfin, en majorité les corrélations entre LISREL et l'approche alternative sont plus grandes, qu'entre LISREL et PLS, pour le modèle ECSI. Bien évidemment, il faut rester prudent sur ces résultats, car ils ne sont valables que pour cette étude.

TABLEAU 4
Coefficients de détermination R^2-ECSI

Variables latentes	**LISREL**	**PLS**	**RFPC**
Attente	0,946	0,162	0,137
Qual. perçue	0,513	0,045	0,035
Val. perçue	0,572	0,205	0,181
Image	0,000	0,000	0,000
Satisfaction	0,762	0,642	0,637
Fidélité	0,289	0,245	0,245

TABLEAU 5
Coefficients de corrélation entre mêmes variables latentes – ECSI

Variables latentes	**LISREL *vs* PLS**	**LISREL *vs* RFPC**	**PLS *vs* RFPC**
Attente	0,982	0,996	0,976
Qual. Perçue	0,980	0,986	0,999
Val. perçue	0,963	0,978	0,994
Image	0,995	0,995	0,999
Satisfaction	0,999	0,996	0,999

4.2. Approche PLS vs *approche RFPC : Une approche géométrique*

4.2.1. Description mathématique

Après avoir analysé les résultats de façon empirique et établi certains liens entre les méthodes, nous allons maintenant nous intéresser aux liens théoriques qui peuvent exister entre l'approche PLS et l'approche alternative.

En effet ces deux méthodes présentent de grandes similitudes : à partir du modèle fixé, chaque variable latente est estimée au niveau des individus. Puis à partir de ces résultats, on applique le modèle linéaire, pour établir les liens entre variables, et les coefficients de régression sont estimés en utilisant la méthode des moindres carrés ordinaires comme le résume le tableau ci-après :

Nous effectuons la comparaison des deux méthodes au niveau de l'étape 1 : composantes PLS *vs* composantes principales. Si celles-ci sont très proches, l'étape 2 consistant, dans les deux cas, à une régression, les résultats finaux (estimations des coefficients) devraient être également très voisins.

TABLEAU 6
Comparaison de l'approche PLS et de l'approche RFPC

	Approche PLS	**Méthode RFPC**
Etape 1	Variables latentes = Premières composantes PLS	Variables latentes = Premières composantes Principales
Etape 2	Régression linéaire ordinaire	Régression linéaire ordinaire

Pour une telle comparaison, nous avons utilisé un article de De Jong *et al.* [DEJ 97].

L'idée de base de l'article est d'étudier le modèle de régression standard et centré défini par :

$$y = X\beta + \varepsilon \tag{28}$$

y est un vecteur $(n \times 1)$, X est une matrice $(n \times p)$, β est le vecteur des paramètres $(p \times 1)$ et ε est le vecteur des erreurs $(n \times 1)$, formé d'observations indépendantes, identiquement distribuées de moyenne 0 et de variance σ^2.

L'objectif est d'estimer le vecteur β et pour ce faire les auteurs proposent trois méthodes : la méthode des moindres carrés ordinaires, la régression sur composantes principales et la régression PLS. Les estimateurs de β pour chacune de ces méthodes seront notés respectivement : $\hat{\beta}_{ols}$, $\hat{\beta}^m_{pcr}$ et $\hat{\beta}^m_{pls}$.

Dans l'espace $\mathrm{I\!R}^n$, $X\hat{\beta}^m_{pcr}$ est la projection de y et donc (d'après le théorème des trois perpendiculaires) de $X\hat{\beta}_{ols}$ sur l'espace engendré par les m premières composantes principales de X. De même, $X\hat{\beta}^m_{pls}$ est la projection de y et donc de $X\hat{\beta}_{ols}$ sur l'espace engendrée par les m premières composantes PLS de X.

Géométriquement, la première composante PLS a d'autant plus de chances d'être proche de la première composante principale que la différence entre la première valeur propre de l'ACP de X et les autres est élevée, alors qu'elle risque d'en être d'autant plus éloignée que toutes les valeurs propres dans l'ACP de X sont voisines (et donc voisines de 1). Dans ce dernier cas, à la limite quand toutes les valeurs propres sont égales à 1 (auquel cas les variables « explicatives » sont non corrélées), $\hat{\beta}^m_{pls} = \hat{\beta}_{ols}$ (avec $m = 1$) et donc $X\hat{\beta}^m_{pls} = X\hat{\beta}_{ols}$, alors que $\hat{\beta}^m_{pcr}$ et $X\hat{\beta}^m_{pcr}$ sont indéterminés (puisque tout vecteur normé du sous-espace engendré par les colonnes de X dans $\mathrm{I\!R}^n$ est une composante principale).

Examinons maintenant de façon précise dans quel cas la première composante PLS de X est proportionnelle à la première composante principale (auquel cas $X\hat{\beta}^1_{pls} = X\hat{\beta}^1_{pcr}$). Supposons que dans l'ACP de X on ait $r(r \leqslant p)$ composantes principales normées, ξ_a non triviales (*i.e.* associées à des valeurs propres λ_a non nulles) rangées par valeurs propres décroissantes. Alors la première composante PLS est proportionnelle à la première composante principale seulement dans les deux cas suivants :

(i) $r = 1$

(ii) $r > 1$, y est orthogonal à ξ_a pour $a = 2, r$

Si (i) ou (ii) est réalisé, on a : $\hat{\beta}^1_{pls} = \hat{\beta}^1_{pcr}$, ainsi que $X\hat{\beta}_{ols} = X\hat{\beta}^1_{pls} = X\hat{\beta}^1_{pcr}$ ($\hat{\beta}_{ols}$ étant indéterminé si $r \neq p$). Si de plus on est dans le cas (ii) avec $r = p$, les trois estimateurs $\hat{\beta}_{ols}$, $\hat{\beta}^1_{pls}$ et $\hat{\beta}^1_{pcr}$ coïncident (*cf.* annexe pour la démonstration).

Dans le cas extrême, et en se plaçant dans le cadre du $j^{\text{ième}}$ bloc de variables, où les variables x_{jh} sont parfaitement corrélées, alors la première valeur propre est égale à p_j (le nombre de variables du bloc j) et les autres valeurs propres sont égales 0. Au paragraphe 3.3, nous avons vu que le poids PLS : $w^*_{jh} = \text{cov}(x^*_{jh}, Z_j)$, où Z_j est une composante PLS centrée-réduite et que le poids RFPC : $\alpha_{1jh} = \text{cov}(x^*_{jh}, Y^*_{1j})/\lambda_{1j}$. Alors, dans ce cas extrême, le vecteur w^*_{jh} est proportionnel au vecteur α_{1jk}, en effet :

$$w^*_{jh} = \text{cov}(x^*_{jh}, Z_j) = \text{cov}(x^*_{jh}, Y^*_{1j}) = p_j \alpha_{1jh} \tag{29}$$

Cette analyse revient à se poser le problème d'unidimensionnalité des blocs de variables manifestes. En effet, quand la première valeur propre associée à un bloc est élevée (très supérieure à l'unité), alors une grande partie de l'inertie sera expliquée par la première composante principale, on pourra donc considérer que ce bloc est unidimensionnel. En d'autres termes, les résultats des deux approches seront comparables. A l'opposé, si la première valeur propre est assez proche de l'unité, alors l'inertie totale se décomposera de façon relativement équitable, d'où la suspicion d'un bloc nettement moins unidimensionnel. Dans ce cas, les résultats entre les méthodes auront tendance à diverger.

Afin de confirmer ces résultats nous avons effectué des simulations (non présentées ici) dans les deux cas extrêmes : celui où la première valeur propre est très importante par rapport aux autres et celui où, au contraire, les valeurs propres ont des valeurs assez proches, puis nous avons comparé les estimations des variables latentes avec l'approche PLS et la méthode alternative. Les résultats obtenus vérifient bien la discussion précédente : quand les variables manifestes sont très corrélées dans chaque variable latente, les premières valeurs propres sont très élevées, et la première composante principale et la composante PLS sont très corrélées, (forte unidimensionnalité du bloc), d'où des résultats pratiquement identiques pour les régressions entre variables latentes. *A contrario*, quand les variables manifestes ne sont pas corrélées du tout, pour chaque variable latente, les premières valeurs propres sont peu différentes des autres valeurs propres, par conséquent les premières composantes principales et les composantes PLS ne sont pas corrélées (ou très peu), les blocs sont moins unidimensionnels, alors les résultats obtenus en régression différent.

4.2.2. Application sur les résultats de l'étude

Les tableaux suivants concernent le modèle ECSI et fournissent les poids relatifs de chaque variable manifeste. Dans la construction de la variable latente. Ils sont accompagnés de la valeur de la première valeur propre.

Comme on l'avait constaté dans le paragraphe 4.1.1., en ce qui concerne la variable « Attente », les poids relatifs sont relativement éloignés. Une des raisons est que la première valeur propre bien qu'assez élevée, n'a qu'une contribution (39,7%) d'un peu plus de deux fois son espérance : $100/6\,\% = 16,7\,\%$. Celle-ci n'est donc pas assez forte pour expliquer la variance des données liées à la variable « Attente »

TABLEAU 7
Comparaison de l'approche PLS et de l'approche RFPC – ECSI Attente

$\lambda_1 = 2,357$ (39,3 % *vs* 16,7 %)

Variables manifestes	**Approche PLS**	**Approche RFPC**
b1	0,234	0,163
b2	0,220	0,127
b3	0,142	0,183
b4	0,087	0,149
b5	0,167	0,189
b6	0,150	0,189

(cf. tableau 7). D'ailleurs, la moyenne des différences absolues entre les poids relatifs est de loin la plus élevée avec 0,055 (*cf.* tableau 2).

La valeur perçue est également relativement mal reconstituée, car la première valeur propre observée représente seulement un peu plus d'une fois et demi son espérance (*cf.* tableau 8). La moyenne des différences absolues de 0,033, arrive juste après (en ordre décroissant) celle liée à la variable « Attente » (*cf.* tableau 2).

TABLEAU 8
Comparaison de l'approche PLS et de l'approche RFPC – ECSI Valeur perçue

$\lambda_1 = 1,621$ (40,5 % *vs* 25,0 %)

Variables manifestes	**Approche PLS**	**Approche RFPC**
co1	0,251	0,222
co2	0,226	0,268
pv1	0,342	0,305
pv2	0,180	0,205

Par contre, la qualité perçue est très bien reconstituée, car la première valeur propre représente à peu près six fois la valeur de son espérance et la moyenne des différences absolues des poids relatifs est très petite : 0,005. On peut constater ce fait par la forte proximité des poids relatifs (*cf.* tableau 9).

De même, la première valeur propre observée pour l'image est de deux fois et demi son espérance. La moyenne des différences absolues des poids relatifs est légèrement plus élevée que celle de la qualité perçue, avec 0,006 (*cf.* tableau 2). On peut d'ailleurs constater que les poids relatifs sont vraiment très proches pour les deux méthodes (*cf.* tableau 10).

Enfin, la satisfaction est moyennement bien reconstituée : la première valeur propre représente un peu plus d'une fois et demi son espérance, d'ailleurs la moyenne

TABLEAU 9
Comparaison de l'approche PLS et de l'approche RFPC – ECSI Qualité perçue

$\lambda_1 = 5,830$ (30,7 % *vs* 5,3 %)

Variables manifestes	**Approche PLS**	**Approche RFPC**
r1	0,063	0,058
r2	0,070	0,062
r3	0,076	0,064
r4	0,046	0,041
c1	0,065	0,055
c2	0,054	0,061
c3	0,057	0,056
f1	0,038	0,049
f2	0,048	0,054
f3	0,048	0,040
f4	0,056	0,057
i1	0,048	0,047
i2	0,062	0,062
i3	0,056	0,055
t1	0,045	0,052
t3	0,029	0,042
t4	0,045	0,046
t5	0,050	0,052
t6	0,043	0,045

TABLEAU 10
Comparaison de l'approche PLS et de l'approche RFPC – ECSI Image

$\lambda_1 = 2,548$ (50,8 % *vs* 20,0 %)

Variables manifestes	**Approche PLS**	**Approche RFPC**
cf1	0,218	0,207
cf2	0,200	0,215
im1	0,227	0,226
im2	0,122	0,118
im3	0,233	0,234

TABLEAU 11
Comparaison de l'approche PLS et de l'approche RFPC – ECSI Satisfaction

$\lambda_1 = 1,613$ (53,4 % *vs* 33,3 %)

Variables manifestes	**Approche PLS**	**Approche RFPC**
sat1	0,406	0,368
sat2	0,284	0,295
sat3	0,310	0,337

des différences absolues des poids relatifs vaut 0,025, et arrive en troisième position (en ordre décroissant) (*cf.* tableaux 2 et 11).

L'approche RFPC proposée ouvre des perspectives très intéressantes et nous a permis d'établir un premier pont théorique entre cette méthode et l'approche PLS.

Car si le modèle conceptuel marketing de départ reflète bien la réalité *i.e.* si les groupes établis dans le schéma structurel décrivent effectivement une même variable latente, alors logiquement les variables manifestes concernées sont liées assez fortement entre elles. Et dans ce cas, on a vu que la méthode alternative, très simple tant au niveau des techniques mathématiques requises qu'au niveau de la mise en œuvre pratique, est aussi efficace que l'approche PLS. Mais ceci n'est valable que si le modèle est précis d'où une raison de plus pour s'interroger sur la validité des modèles utilisés.

Par conséquent, le problème de cohérence du modèle conceptuel nous a poussé à tenter d'établir un test statistique afin de valider le modèle ECSI, mais également de construire notre propre modèle.

5. Construction d'un modèle « libre »

Toute l'étude faite jusqu'à présent repose sur des hypothèses de travail très fortes à savoir le modèle conceptuel ECSI dont nous nous sommes servis et qui est figé au moins à trois niveaux :

– dans le choix des variables manifestes correspondant aux questions de l'enquête,

– dans la détermination des groupes de variables manifestes constituant les variables latentes,

– dans la présence ou l'absence de liens entre les groupes, ainsi que le sens de ces liaisons.

Si l'on considère que le premier niveau reste imposé, les deux autres peuvent en revanche être remis en question.

5.1. Proposition de tests théoriques

5.1.1. Premier test

A partir du tableau initial des réponses, on peut soit établir une typologie des variables manifestes à l'aide d'une classification hiérarchique descendante, permettant de rechercher des composantes obliques [3] . On obtient ainsi plusieurs classes qui regroupent les différentes variables (cf. §5.1.2.); soit utiliser tel quel les groupes de variables manifestes déterminées par les experts en marketing. Ensuite, nous construisons un modèle de régression multiple de la façon suivante :

Supposons que la classification ou la typologie des experts possède R classes et que $Y_1, \ldots, Y_R$, soient les variables latentes associées à ces classes. Soit Y_j une de ces variables, les $R-1$ autres sont les variables candidates à l'explication de Y_j. L'approche corrélation semble assez indiquée afin d'établir quelles variables jouent effectivement un rôle dans l'explication de Y_j.

Cependant, il arrive fréquemment que la dépendance apparente entre deux variables soit due en réalité aux variations d'une troisième variable, nous avons donc choisi d'utiliser les coefficients de corrélation partielle afin d'éliminer cette influence d'une ou plusieurs variables. Il suffira ensuite de déterminer un seuil à partir duquel on décide qu'il existe ou non un lien entre deux variables, à l'aide du test statistique classique sur le coefficient de corrélation partielle.

Il y a $R(R-1)/2$ coefficients à calculer, car les corrélations étant symétriques si deux variables ont un lien, on pourra dire qu'elles s'influencent l'une et l'autre sans toutefois préciser le sens de la liaison *i.e.* la variable explicative et celle expliquée. Cependant, en analysant les questions regroupées dans les deux variables latentes concernées, ce sens sera peut-être intuitif.

Par cette analyse *a posteriori*, on obtient un nouveau modèle, que l'on appelle le modèle de référence, et que l'on doit valider statistiquement.

Pour cela, on tire B échantillons Bootstrap (tirage avec remise directement sur les données) et pour chacun d'eux on calcule les liens entre les variables latentes comme précédemment. On obtient ainsi B modèles conceptuels libres dont, on l'espère, le modèle de référence fait partie. Il faut ensuite tester :

H_0 : « le modèle de référence est apparu au hasard » contre

H_1 : « le modèle de référence n'est pas apparu au hasard »

Le nombre total de différents modèles possibles (à partir de groupes fixés) est $2^{R(R-1)/2}$ car on a $R(R-1)/2$ liens possibles et le lien peut exister ou pas.

Sous l'hypothèse H_0, la loi d'apparition du modèle de référence est donc une loi de Bernoulli de paramètre p avec p égal à l'inverse de $2^{R(R-1)/2}$.

On pose $Z_i = 1$ ou 0 suivant qu'au i$^{\text{ème}}$ tirage Bootstrap le modèle apparu est le modèle de référence ou non, ainsi sous H_0, $Z_i \sim > \mathcal{B}(1,p)$, la loi de Bernoulli de paramètre p.

---

[3] Nous avons utilisé la procédure VARCLUS du Système SAS [SAS 01].

Soit $M = \sum_{i=1}^{B} Z_i$ *i.e.* le nombre de fois où le modèle de référence est apparu, alors sous H_0, M suit une loi Binomiale : $\mathcal{B}(B, p)$. Soit m le nombre effectivement observé d'apparition du modèle. Il suffit ensuite de comparer m au fractile m_0 avec un risque de première espèce de 5 %, par exemple. Si $m > m_0$, alors on rejette H_0, sinon on la garde.

5.1.2. Second test

A partir des données on fait une classification des variables puis on cherche à tester si cette classification, que l'on appellera classification de référence, est apparue au hasard ou non. On désignera par k le nombre de classes de cette classification.

On teste donc : H_0 : « les groupes sont apparus au hasard » contre

H_1 : « les groupes ne sont pas apparus au hasard »

Pour cela, on tire de nouveau, B échantillons Bootstrap et à chacun on construit une typologie de variables qui propose des groupes que l'on compare aux groupes de référence.

Soit P le nombre total de classifications possibles *i.e.* le nombre de partitions en k classes d'un ensemble à n éléments, alors P est le nombre de Stirling de deuxième espèce noté $P(n, k)$ défini pour $k \leqslant n$ et qui satisfait à l'équation de récurrence suivante :

$$P(n,k) = P(n-1,k-1) + kP(n-1,k)$$
$$\text{avec } P(n,1) = P(n,n) = 1 \text{ et } P(n,n-1) = n(n-1)/2.$$

Soit $Z_i = 1$ ou 0 suivant qu'au i$^{\text{ème}}$ tirage Bootstrap la classification obtenue est celle de référence ou non. Si p est l'inverse de $P(n, k)$ alors sous H_0, $Z_i \sim> \mathcal{B}(1, p)$. Comme précédemment, on détermine ensuite le nombre de fois où on a obtenu la même classification que celle de référence : $M = \sum_{i=1}^{B} Z_i$ et sous l'hypothèse H_0, $M \sim> \mathcal{B}(B, p)$. Soit m le nombre effectivement observé d'apparition du modèle. Il suffit ensuite de comparer m au fractile m_0 avec un risque de première espèce de 5 %, par exemple. Si $m > m_0$, alors on rejette H_0, sinon on la garde.

5.2. Tests effectués

5.2.1. Validité du modèle ECSI

A partir des six groupes définis par le modèle ECSI, nous avons établi nos propres liens grâce aux corrélations partielles, ceci nous donne le modèle 1. Il comporte 8 liens, alors que ECSI en avait 10. Parmi cela, seuls 5 sont communs (*cf.* figure 7).

Nous effectuons les deux tests :

H_0 : « le modèle 1 est apparu au hasard » contre

H_1 : « le modèle 1 n'est pas apparu au hasard »

et :

H'_0 : « le modèle ECSI est apparu au hasard » contre

H'_1 : « le modèle ECSI n'est pas apparu au hasard »

Nous tirons 1000 échantillons Bootstrap et nous notons m le nombre de fois où apparaît le modèle 1 et m' le nombre d'apparitions du modèle ECSI. Sous H_0 (resp H'_0), m (resp m') est censé être une réalisation d'une variable aléatoire Binomiale de paramètre B et p avec ici $B = 1000$ et $p = 1/2^{R(R-1)/2}$. Comme il y a 6 groupes, p vaut $1/2^{15}$.

On obtient $m = 282$ et $m' = 0$.

La probabilité p de la binomiale étant très petite, il est clair que le premier résultat nous permet de rejeter (presque à 100 %) l'hypothèse H_0. Par conséquent, le modèle 1 n'est pas apparu au hasard. Le second résultat ne nous permet cependant pas de conclure.

On se rend vite compte que ce problème se posera pour les deux tests envisagés ci-dessus.

Cependant, les fractiles pour un risque à 5 % étant très faibles, ces tests permettent de conclure quand le nombre d'apparitions de tel ou tel modèle dépasse 2 ou 3.

5.2.2. Classifications propres et comparaison à celles d'origine

Afin de voir, *a posteriori*, si les différents groupes du modèle ECSI étaient raisonnables nous avons procédé à des classifications des variables en 6 groupes. La classification en 6 groupes est assez loin de la classification donnée par le modèle ECSI. En effet, à part la variable latente « besoin » (attente) qui regroupait b1, b2, b3, b4, b5 et b6 et que l'on retrouve dans une même classe, les autres groupes ne se retrouvent pas de façon précise. Il semble que le modèle ECSI ne soit pas ici très approprié pour le questionnaire.

5.3. Recherche de modèles marketing

5.3.1. Premier modèle

Tout d'abord, on décide de garder les six mêmes groupes que ceux établis afin d'étudier le modèle ECSI *i.e.* les groupes fournis dans le tableau 1. On établit nos propres liens grâce aux corrélations partielles (cf. §5.1.1.).

Nous trouvons 8 liens entre les variables, dont 5 en commun avec le modèle ECSI (qui en comportait 10 au total) comme le montre la figure 6.

Les corrélations partielles ne nous permettent pas, *a priori*, de donner un sens aux différents liens, cependant avec l'aide d'experts en marketing, nous avons pu établir ces liens assez logiquement. Cela donne le modèle de la figure 7.

On remarque qu'il n'y a pas de liens entre la satisfaction et la fidélité ce qui remet en question le fait que les personnes satisfaites ne sont pas forcément fidèles et inversement. Au contraire le coefficient entre la fidélité et l'image était très significatif et on retrouve le lien correspondant. Ceci prouve bien l'importance de l'image sur

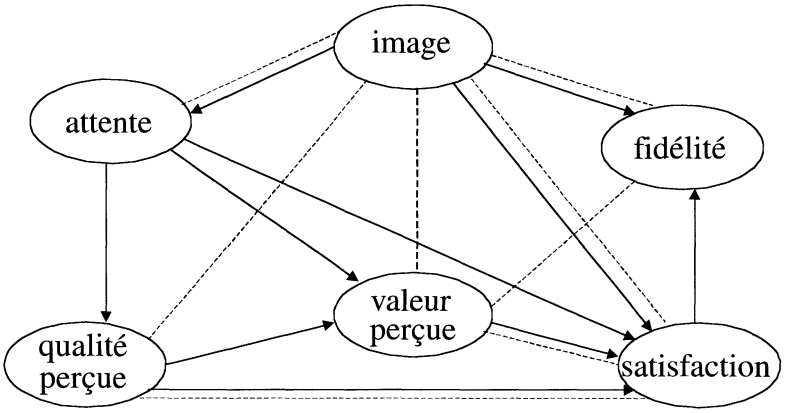

FIGURE 6
Modèle ECSI et nouveau modèle

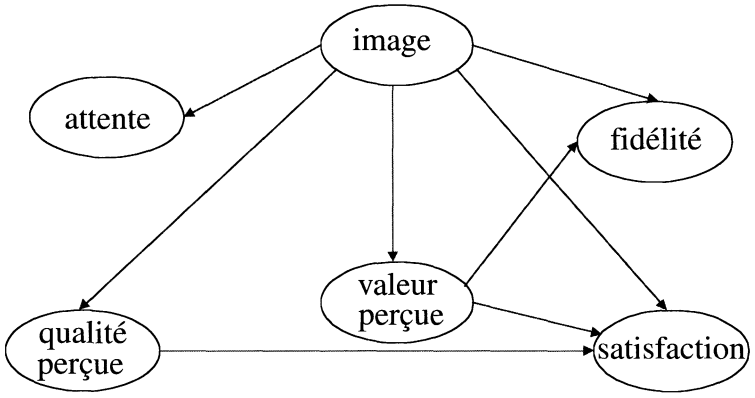

FIGURE 7
Premier modèle

la fidélité par rapport à la satisfaction contrairement à ce que l'on pourrait penser de prime abord.

5.3.2. Second modèle

Pour ce deuxième modèle (cf. §5.2.2.), on ne se fixe aucun *a priori* à part le nombre de groupes fixé à 6 pour construire la typologie de variables, afin de la comparer aux groupes ECSI.

Le pourcentage de variance expliquée est de 45 % et les six classes sont organisées de la façon suivante :

Classe 1 : i1, i2, i3

Classe 2 : t1, t3, t4, t5, t6, sat3

Classe 3 : pv2, b1, b2, b3, b4, b5, b6

Classe 4 : co2, pv1, cf2, im1, im3, fid1, sat2

Classe 5 : r4, co1, im2

Classe 6 : r1, r2 , r3, c1, c2 , c3, f1, f2, f3, f4, cf1, sat1

La classe 5 regroupe peu de variables et n'est guère significative (d'autres études montrent que les items qui lui sont liés sont assez peu représentatifs), on décide par conséquent de la supprimer.

Après une étude des différentes questions à l'intérieur de chaque classe, on décide de donner les noms suivants aux 5 groupes restants :

Classe 1 : information

Classe 2 : qualité des produits/ relation technique

Classe 3 : besoins futurs

Classe 4 : attachement des clients

Classe 6 : facturation/ relation clientèle

On calcule ensuite les corrélations partielles de ces 5 groupes afin d'observer les liens entre les groupes et toujours avec l'appui de spécialistes marketing, on détermine le sens des flèches. Cela nous conduit au modèle de la figure 8 :

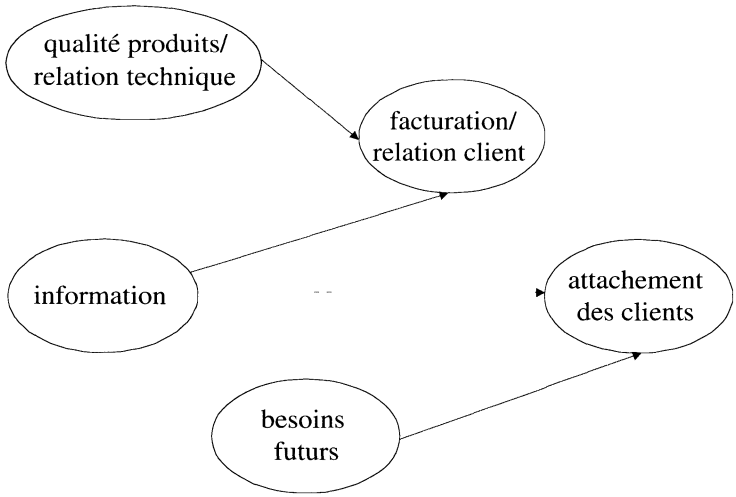

FIGURE 8
Second modèle

La variable à expliquer est ici l'attachement des clients à l'entreprise qui regroupe la fidélité et la satisfaction des modèles précédents.

Le modèle ECSI, longtemps pris comme modèle de référence, semble être de plus en plus remis en question. On a d'ailleurs pu remarquer, sur notre jeu de données, le peu d'influence que semblait avoir la satisfaction sur la fidélité. D'autres modèles sont donc à envisager. On peut établir autant de classifications et de modèles que d'études réalisées. Cependant on peut espérer qu'un modèle particulier aura tendance à se singulariser et que celui-ci pourrait être le plus logique possible au niveau de l'interprétation.

6. Apports, critiques et propositions

La principale différence entre l'approche PLS et la méthode RFPC d'une part, et l'approche LISREL d'autre part réside dans le mode d'estimation du schéma structurel. L'estimation du modèle par l'approche PLS tout comme la méthode alternative se fait en deux temps : tout d'abord l'estimation du modèle externe qui relie les variables latentes aux variables manifestes, puis celle du modèle interne grâce à des régressions multiples. Il faut tout de même noter que l'estimation des variables latentes par PLS est faite de façon itérative et en tenant compte du modèle interne (même si celui-ci ne semble pas jouer un rôle capital dans l'estimation), ce qui n'est pas le cas de la méthode alternative où les variables latentes sont estimées directement et de la même façon quel que soit le modèle interne. Mais nous avons pu montrer géométriquement que dans le cas de groupes bien représentés par leur première valeur propre, ces estimations convergeaient. Cependant, si l'on se place dans une optique de prévision, alors PLS est meilleure que RFPC, car les composantes PLS dépendent par construction de la liaison entre la variable à expliquer et les variables explicatives, ce qui n'est pas le cas de l'approche alternative. On peut alors se placer sous deux angles différents. Pour l'estimation du modèle, PLS force les liaisons (même si elles sont faibles), alors que ce n'est pas le cas de RFPC. Alors, l'approche PLS, bien que biaisée, semble meilleure que l'approche RFPC, mais si le modèle proposé n'est pas correct, alors cette dernière méthode, du fait qu'elle soit non biaisée, sera meilleure. Par conséquent, l'extrapolation (ou la prédiction) avec PLS peut être dangereuse, d'où l'obligation dans ce cas de faire de la validation croisée, ce qui ne sera pas utile pour RFPC. Cet aspect figé et contraint du modèle est encore plus présent dans LISREL, car l'estimation du schéma structurel (*i.e.* les modèles internes et externes) est réalisée de façon globale. Cette procédure est très utilisée, cependant elle ne permet pas d'estimer tous les modèles, certains ne sont pas identifiables. Il faut alors changer de modèle ou de méthode.

Dans un cadre où le modèle conceptuel est proche de la réalité et adapté aux données dont on dispose, les méthodes PLS et alternative sont très voisines et l'on peut même dire équivalentes. L'avantage principal de la méthode RFPC sur l'approche PLS est que les outils mathématiques utilisés sont très simples et parfaitement maîtrisés contrairement à l'approche PLS qui repose sur un principe d'estimation itérative, et que l'on pourrait plutôt qualifier de « boite noire ». D'autre part, nous avons souligné l'importance du choix du modèle au départ, ce qui nous a conduit à en proposer de nouveaux. Cela nous a naturellement amenés ensuite à nous poser la question de l'adéquation du modèle aux données.

Il reste cependant encore deux problèmes à résoudre. Le premier a trait à la mise en place d'un indicateur de qualité globale du modèle afin de pouvoir en comparer plusieurs et d'en tirer le « meilleur ». Ces travaux sont pour l'instant encore à l'état de recherche, mais quelques pistes commencent à voir le jour. Le second problème concerne le sens des liaisons quand on a établi un modèle à l'aide des corrélations partielles. Là aussi, nous avons quelques idées mais nous ne les présentons pas ici, car les recherches ne sont pas complètement abouties.

Enfin, il est tout de même important, de rappeler que quelle que soit l'approche adoptée pour établir un modèle conceptuel ou pour en estimer un fixé a priori, le questionnaire, notamment la formulation des questions et la qualité des données recueillies sont primordiales pour obtenir des résultats que l'on est en droit d'interpréter. Sinon, l'application de méthodes les plus sophistiquées ou celles dont on pense qu'elles sont les plus adéquates pour résoudre le problème, n'apporte aucune information supplémentaire et surtout valide.

Remerciements. – Nous tenons à remercier Pierre Cazes et Michel Tenenhaus pour leurs observations, suggestions et appui documentaire.

Bibliographie

[ASL 00] Association Léonard de Vinci, Mouvement Français pour la qualité, *Présentation du projet ECSI*, 2000.

[AUR 00] AURIER P., Validité des composantes et relation à la marque, *Extrait du 17ème Congrès International de l'Association Française de Marketing*, 2000.

[BAR 87] BARTHOLOMEW J., *Latent Variable Models and Factor Analysis*, Oxford University Press, 1987.

[BAY 00] BAYOL M.-P., DE LA FOYE A., TELLIER A. & TENENHAUS M., Use of PLS Path Modelling to estimate the European Consumer Satisfaction Index (ECSI) model, *Statistica Applicata*, Vol. 12, n° 3, 2000.

[BOL 89] BOLLEN K. A., *Structural Equations with Latent Variables*, Wiley-Interscience, 1989.

[CHI 98] CHIN W.W., The Partial Least Squares Approach for Structural Equation Modelling, in G.A. Marcoulides (Ed.), *Modern Methods for Business Research*, Lawrence Erlbaurm Associates, pp. 295-336.

[CHI 99] CHIN W.W. & NEWSTED P.R., Structural Equation Modelling Analysis with Small Sample using Partial Least Squares, in R. Höyle (Ed.), *Statistical Strategies for Small Sample Research*, Sage Publications, pp. 307-341.

[CHI 01] CHIN W.W., *PLS-Graph User's Guide*, C.T. Bauer College of Business, University of Houston, USA.

[DEJ 97] DE JONG S. & PHATAK A., The geometry of partial least squares, *Journal of Chemometrics*, Vol. 11, 312-318, 1997.

[DER 97] DERQUENNE C., Le Bootstrap : un outil de calcul pour la statistique appliquée, *rapport technique EDF/DER*, 1997.

[DER 00] DERQUENNE C., Critiques de l'approche PLS et proposition d'une méthode alternatives, document interne, EDF R& D, 2000.

[DER 02a] DERQUENNE C. & HALLAIS C., Une méthode alternative à l'approche PLS : Comparaison et application aux modèles conceptuels marketing, Actes des Journées de Statistique, Bruxelles, 2002.

[DER 02b] DERQUENNE C. & HALLAIS C., Une méthode alternative à l'approche PLS : Comparaison et application aux modèles conceptuels marketing, 1er Colloque Franco-Libanais sur la Statistique et l'Analyse des Données dans les Sciences Appliquées et Economiques, Beyrouth, 2002.

[FOR 92] FORNELL C., A National Customer Satisfaction Barometer : The Swedish Experience, *Journal of Marketing*, Vol. 56, 6-21, 1992.

[FOR 94] FORNELL C. & CHA J., Partial Least Squares, in *Advanced Methods of Marketing Research*, R.P. Bagozzi (Ed.), Basil Blackwell, Cambridge, MA., pp. 52-78, 1994.

[JOR 79] JÖRESKOG K.G. & SÖRBUM D., *Advance in Factor Analysis and Structural Equation Models*, Abt Books, Cambridge, 1979.

[JOR 82] JÖRESKOG K.G. & WOLD H., The ML and PLS techniques for modeling with latent variables : Historical and comparative aspects in *System under indirect observation*, vol 1, Jöreskog K.G. & Sörbum D. (Eds), North-Holland, Amsterdam, pp. 263-270, 1982.

[LOH 87] LOHMÖLLER J.-B., *LVPLS Program Manual, Version 1.8*, Zentralarchiv für Empirische Sozialforschung, Köln, 1987.

[LOH 89] LOHMÖLLER J.-B., *Latent Variables Path Modeling with Partial Least Squares*, Physica-Verlag, Heildelberg, 1989.

[MAR 89] MARTENS H. & NÆS T., *Multivariate Calibration*. John Wiley & Sons, New York, 1989.

[MUE 96] MUELLER R.O., *Basic Principles of Structural Equations Modelling : An Introduction to LISREL and EQS*, Springer-Verlag, New York, 1996.

[SAS 01] SAS Institute Inc., *What's New in SAS*, Release 8.1 and 8.2, Online Documentation, 2001.

[TEN 98] TENENHAUS M., *La Régression PLS*, Éditions Technip, Paris, 1998.

[TEN 99] TENENHAUS M., L'approche PLS, *Revue de Statistique Appliquée*, Vol. 47, n° 2, pp. 5- 40, 1999.

[TEN 01] TENENHAUS M. & GONZALEZ P.-L., Comparaison entre les approches PLS et LISREL en modélisation d'équations structurelles : Application à la mesure de la satisfaction clientèle, *Compte rendu du Club SAS 2001.*

[WOL 82] WOLD H., Soft modeling : the basic design and some extensions, in *System under indirect observation*, vol. 2, Jöreskog K.G & Sörbum D. (Eds), North-Holland, Amsterdam, pp. 1-54, 1982.

[WOL 85] WOLD H., Partial Least Squares, in *Encyclopedia of Statistical Sciences*, vol. 6, Kotz, S. & Johnson, N.L. (Eds), John Wiley & Sons, New York, pp. 581-591, 1985.

Annexe : Démonstration à propos des trois estimateurs : $\hat{\beta}_{pls}$, $\hat{\beta}_{pcr}$, $\hat{\beta}_{ols}$ qui coïncident

Désignons respectivement par λ_α, $\underline{u}_\alpha$, $\underline{z}_\alpha = X\underline{u}_\alpha = \sqrt{\lambda_\alpha}\xi_\alpha$ pour $\alpha = 1, p$, les valeurs propres rangées en ordre décroissant de valeurs (et supposées toutes non nulles, dans un premier temps), les vecteurs axiaux factoriels et les composantes principales issus de l'ACP de X.

Si dans le sous-espace de $|\mathrm{R}^n$ engendré par les colonnes de X, on prend la base formée par les composantes principales $\underline{z}_\alpha$, la régression de $\underline{y}$ est immédiate puisque les $\underline{z}_\alpha$ sont orthogonaux, le $\alpha^{ième}$ coefficient de régression $\hat{\gamma}_\alpha$ étant égal à $\mathrm{cov}(\underline{z}_\alpha, \underline{y})/\lambda_\alpha$, avec $\lambda_\alpha = \mathrm{var}(\underline{z}_\alpha)$. On a alors :

$$\hat{y}_{ols} = \sum_{\alpha=1}^{p} \hat{\gamma}_\alpha \underline{t}_\alpha = X \sum_{\alpha=1}^{p} \hat{\gamma}_\alpha \underline{u}_\alpha = X\hat{\beta}_{ols} \tag{30}$$

d'où on en déduit (puisque X est injective, les λ_α étant tous non nuls) que :

$$\hat{\beta}_{ols} = \sum_{\alpha=1}^{p} \hat{\gamma}_\alpha \underline{u}_\alpha = \sum_{\alpha=1}^{p} \frac{\mathrm{cov}(\underline{z}_\alpha, \underline{y})}{\lambda_\alpha} \underline{u}_\alpha \tag{31}$$

Pour l'estimation sur composantes principales, si on garde m composantes, il suffit dans les formules (30) et (31) de restreindre la somme à ses m premiers termes :

$$\hat{\beta}^m_{pcr} = \sum_{\alpha=1}^{p} \frac{\mathrm{cov}(\underline{z}_\alpha, \underline{y})}{\lambda_\alpha} \underline{u}_\alpha \tag{32}$$

$$\hat{y}^m_{pcr} = X\hat{\beta}^m_{pcr} = \sum_{\alpha=1}^{p} \frac{\mathrm{cov}(\underline{z}_\alpha, \underline{y})}{\lambda_\alpha} \underline{z}_\alpha \tag{33}$$

Remarque : Les expressions précédentes pour $\hat{\beta}_{ols}$ et $X\hat{\beta}_{ols}$ supposent que toutes les valeurs propres sont différentes de 0. Si ce n'est pas le cas $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \lambda_r > \lambda_{r+1} = \lambda_{r+2} = \ldots = \lambda_p = 0$, alors il faut se restreindre à l'espace des r premiers axes factoriels ou des r premières composantes principales. Dans ce cas, $\hat{\beta}_{ols}$ n'est plus unique, et une des expressions de $\hat{\beta}_{ols}$ est $\hat{\beta}^r_{pcr}$. Par contre, $X\hat{\beta}_{ols}$ est toujours unique et est égal à $X\hat{\beta}^r_{pcr}$.

Voyons maintenant le cas particulier où $m = 1$.

On désigne par $\underline{t}_1$, la première composante PLS : $\underline{t}_1 = XX'\underline{y}/\|X'\underline{y}\|$

Or d'après la formule de reconstitution des données $\underset{(n,p)}{X} = \sum_{\alpha=1}^{p} \underset{(n,1)}{\underline{z}_\alpha} \underset{(1,p)}{(\underline{u}_\alpha)'}$, d'où $X'y = \sum_{\alpha=1}^{p} (\underline{z}'_\alpha, \underline{y})\underline{u}_\alpha$ et donc :

$$\underline{t}_1 = \sum_{\alpha=1}^{p} (\underline{z}'_\alpha, \underline{y})\underline{z}_\alpha \Big/ \left[\sum_{\alpha=1}^{p} (\underline{z}'_\alpha, \underline{y})^2\right]^{1/2} \tag{34}$$

Nous avons alors :

$$\hat{y}^1_{pls} = X\hat{\beta}_{pls} = \frac{\underline{t}'_1\underline{y}}{\underline{t}'_1\underline{t}_1}\underline{t}_1 = \frac{\underline{t}'_1\underline{y}}{\underline{t}'_1\underline{t}_1\|X'\underline{y}\|} XX'\underline{y} = cXX'\underline{y} \tag{35}$$

$$\text{avec } c = \frac{\underline{t}'_1\underline{y}}{\underline{t}'_1\underline{t}_1\|X'\underline{y}\|} \tag{36}$$

Par conséquent, on en déduit que :

$$\hat{\beta}^1_{pls} = cX'\underline{y} = c\sum_{\alpha=1}^{p} (\underline{z}'_\alpha, \underline{y})\underline{u}_\alpha \tag{37}$$

$$\hat{y}^1_{pls} = X\hat{\beta}^1_{pls} = c\sum_{\alpha=1}^{p} (\underline{z}'_\alpha, \underline{y})\underline{z}_\alpha\, y \tag{38}$$

$$\text{alors que } \hat{\beta}^1_{pcr} = \frac{\text{cov}(\underline{z}_1, \underline{y})}{\lambda_1}\underline{u}_1 \tag{39}$$

Donc $\hat{\beta}^1_{pcr}$ et $\hat{\beta}^1_{pls}$ sont proportionnels si et seulement si :

$$\forall\, \alpha = 2, \ldots, p : \underline{z}'_\alpha\underline{y} = 0 \tag{40}$$

ce qui est réalisé, soit si $\underline{z}_\alpha = 0$ (*i.e.* $\lambda_\alpha = \mathbf{0}$), soit si $\underline{y}$ est non corrélé à $\underline{z}_\alpha$, ce qui permet d'avancer les deux cas suivants :

a) $r = 1$

b) $2 \leqslant r \leqslant p : \forall\, \alpha = 2, r$; $\underline{z}'_\alpha\underline{y} = 0$ ($\underline{y}$ est non corrélé à $\underline{z}_2, \ldots, \underline{z}_r$)

Quand (40) est réalisé, on déduit de (34) que $\underline{t}_1 = \varepsilon\underline{z}_1$, où $\varepsilon = 1$ si $\underline{z}'_1\underline{y} > 0$ et -1 si $\underline{z}'_1\underline{y} < 0$. Il en résulte immédiatement que $\hat{\beta}^1_{pls} = \hat{\beta}^1_{pcr}$, ce que l'on peut vérifier à partir de (36), (37) et (39) la constante c étant alors égale à $1/(n\lambda_1)$ (puisque $(1/n)\underline{t}'_1\underline{t}_1 = (1/n)\underline{z}'_1\underline{z}_1 = \lambda_1$), tandis que $\text{cov}(\underline{z}_1, \underline{y}) = (1/n)\underline{z}'_1\underline{y}$.

Si de plus, $r = p$, les trois estimateurs $\hat{\beta}_{ols}$, $\hat{\beta}^1_{pls}$ et $\hat{\beta}^1_{pcr}$ sont égaux. Par contre, si $r < p$, alors $\hat{\beta}_{ols}$ est indéterminé, mais les trois estimateurs $X\hat{\beta}_{ols}$, $X\hat{\beta}^1_{pls}$ et $X\hat{\beta}^1_{pcr}$ sont identiques.

Remarques :

1) De façon plus générale, si : $\underline{z}'_\alpha \underline{y} = 0$, pour $\alpha = m+1, \ldots, p$, on a : $\hat{\beta}^m_{pcr} = \hat{\beta}^m_{pls} (= \hat{\beta}_{ols}$ si $r = p)$ et : $X\hat{\beta}^m_{pcr} = X\hat{\beta}^m_{pls} = X\hat{\beta}_{ols}$.

2) les formules donnant $\underline{t}_1$, $\hat{\beta}^1_{pls}$ et $X\hat{\beta}^1_{pls}$ supposent $X'\underline{y} \neq 0$. Si $X'\underline{y} = 0$, $\underline{y}$ est orthogonal au sous-espace engendré par les colonnes de X et donc $X\hat{\beta}_{ols} = X\hat{\beta}^1_{pls} = X\hat{\beta}^1_{pcr} = 0$, ce qui représente un cas sans intérêt.