DES COMPORTEMENTS DE NAVIGATION AUX COMPORTEMENTS DE CHOIX SUR INTERNET

Fabrice Le Guel[*]

Université de Rennes 1

CREREG - CNRS

[*] http://perso.univ-rennes1.fr/fabrice.le-guel/ . Membre du Groupe d'analyse de l'Internet et des TIC (GRANITIC), http://www.granitic.univ-rennes1.fr

Des comportements de navigation aux comportements de choix sur Internet

Résumé

L'article se charge de promouvoir l'utilisation des données de navigation pour analyser les comportements des internautes. Un modèle Logit binomial sur données de panel est proposé comme outil d'analyse générique. Le modèle mesure le niveau d'addiction des internautes sur le moteur de recherche Yahoo!

Mots clés : Internet, Modèles de choix discrets, Logit, persistance, données de navigation.

Abstract

The article takes care to promote the use of clickstream data to analyze the behavior of the Internet users. A panel data binomial Logit model is proposed as tool of generic analysis. The model measures the addiction level of the Internet users on the search engine Yahoo!

Keywords : Internet, discrete choice models, Logit, addiction, clickstream data.

JEL: M31, C25.

A l'instar de Solow qui voyait des ordinateurs partout sauf dans les statistiques, il est de coutume d'admettre que l'on voit peu de commerce électronique sauf dans les statistiques. Comment peut-on donc étudier les usages sur Internet, alors que d'une part, trop peu d'individus ont adopté cette technologie en France et que d'autre part, les enquêtes réalisées par les cabinets de conseil ou autres compagnies d'assurance n'utilisent que rarement une méthodologie rigoureuse (Brousseau, 2000) ? Bien souvent, l'optimisme de leurs prévisions va de pair avec leur part de marché respective…

En se tournant davantage vers le monde de l'Economie et du Marketing, nous allons voir que les outils traditionnels de l'économétrie peuvent être utiles à l'analyse des comportements de choix en ligne.

Un résultat majeur de ces recherches concerne l'observation d'une persistance dans les choix des internautes. En d'autres termes, il semble que les individus reviennent quotidiennement sur certains sites. Ce phénomène peut paraître paradoxal aux yeux des hypothèses originelles consenties aux marchés numériques, supposés alors sans friction.

L'existence d'un phénomène de persistance est selon nous crucial à deux niveaux. Du point de vue du marché, les phénomènes de persistance peuvent être à l'origine d'un pouvoir concurrentiel décisif (Klemperer, 1995). Dès lors, une part de la dispersion des prix sur Internet (Pénard, 2002) pourrait s'expliquer par la capacité de certains sites à connaître leurs clients (Lynch et al., 2000). Du point de vue des entreprises, une stratégie de baisse des prix peut amener de nouveaux clients, mais occasionner de fortes pertes si une grande partie des visiteurs sont déjà fidèles au site. Un arbitrage doit donc s'effectuer, et ce dernier dépendra d'une connaissance fine des comportements de chaque visiteur. Cela est devenu possible grâce à l'utilisation des données de navigation.

L'article proposé ici a donc deux objectifs : d'une part, présenter ce nouveau type de données ainsi que les faits saillants de la littérature qui s'y rapporte (section I), d'autre part, tester l'existence d'une persistance à partir des données de navigation d'un échantillon d'internautes (section II).

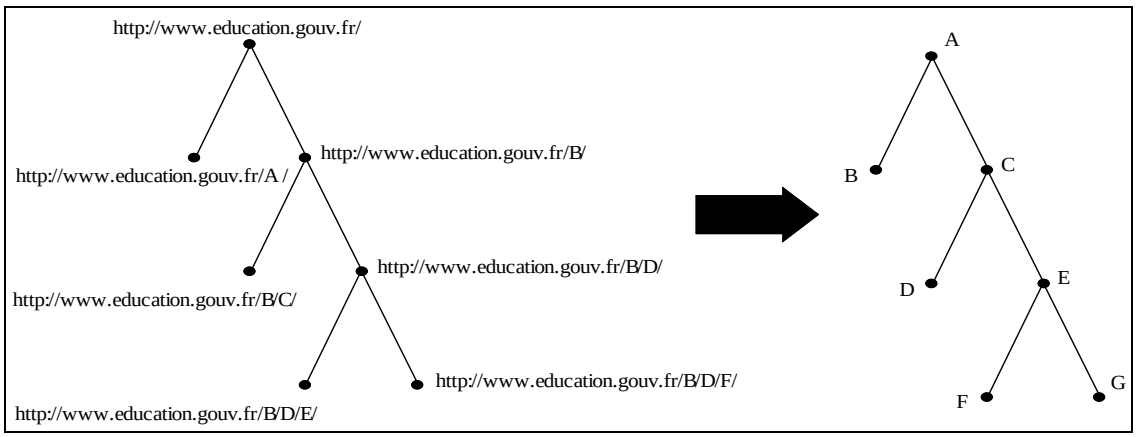
I. Faits saillants sur les comportements des internautes.

1.1/ Définition des données de navigation

Deux types de données peuvent être utilisées lors de l'analyse des comportements de choix des internautes : les données de navigation intra-site, et les données de navigation inter-sites,

toutes deux assimilables aux ‘clickstream data’. Pour mieux comprendre la provenance de ces informations, il est utile de présenter la structure d’un site Internet (figure 1).

Figure 1 : Exemple d'une strucutre simple de site



Les pages d’un site sont matérialisées par une adresse Internet spécifique, appelée Uniform Resource Locator. Ces URL peuvent être plus communément assimilées aux adresses des sites Internet (exemple d’une adresse racine[1] : http://www.education.gouv.fr). Puisque la plupart des sites possèdent plus d’une page, chaque page supplémentaire détient sa propre adresse URL (exemple : http://www.education.gouv.fr/B/). Dès lors, la racine de l’arbre correspond à la page d’accueil du site, chaque point (ou nœud) présente l’adresse d’une page particulière, et les segments reliant ces points annoncent qu’il existe un lien hypertexte amenant aux sous-branches immédiates de l’arbre.

Si nous remplaçons l’adresse URL de chaque page par des lettres (exemple : A pour http://www.education.gouv.fr), le chemin de navigation possible d’un visiteur peut être la séquence ACEF. Autrement dit, l’internaute est arrivé sur la page A du site, puis à cliqué sur un lien pour aller sur la page C du site, et de lien en lien, ce visiteur a terminé sa navigation à la page F.

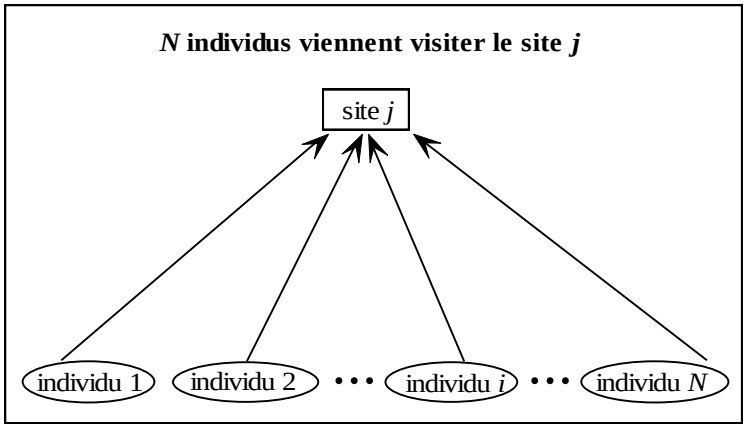
Dans ce sens, les données de navigation correspondent simplement à l’enregistrement de l’adresse URL de chaque page visitée. Plus précisément, les données de navigation sont des fichiers informatiques contenant uniquement du texte. Sous certaines conditions techniques

[1] Appelée ‘racine’ puisqu’elle correspond à la racine de l’arbre présenté dans la figure 1.

(utilisation de 'cookies[2]'), ces fichiers peuvent retracer l'histoire entière des visites d'un internaute sur un site.

Jusqu'ici, nous avons appréhendé les comportements de N individus *sur* un site j (figure 2).

Figure 2 : Données Log au niveau du site *j*



Puisque la navigation sur Internet consiste le plus souvent à aller d'un site à un autre, une acception plus large des comportements de navigation considère I individus face à J sites différents. Sous certaines conditions techniques, ces informations sont là encore enregistrées dans des fichiers texte (encadré 1).

Encadré 1: Données de navigation pour un individu *i*

```
cs17 787125748 56906 "http://www.ludvigsen.dhhalden.no/webdoc/ahlan.gif" 1700 2.691705
cs17 787125764 828867 "http://www.ludvigsen.dhhalden.no/webdoc/ahlan.gif" 0 0.0
cs17 787125785 497644 "ftp://ftp.hiof.no/pub/levant/lebanon/jbeil/jbeil01.jpg" 21839 15.229143
cs17 787126872 204994 "http://www.ludvigsen.dhhalden.no/webdoc/levant_servers.html#Lebanon" 0 0.0
cs17 787126872 389659 "http://www.ludvigsen.dhhalden.no/webdoc/ahlan.gif" 0 0.0
cs17 787128239 578430 "ftp://ftp.hiof.no/pub/levant/music_and_song/music.list" 40418 8.752783
cs17 787128301 795825 "http://www.ludvigsen.dhhalden.no/webdoc/levant_servers.html#Lebanon" 0 0.0
cs17 787128302 100752 "http://www.ludvigsen.dhhalden.no/webdoc/ahlan.gif" 0 0.0
cs17 787128330 9639 "ftp://ftp.hiof.no/pub/levant/music_and_song/singers/AbdelHalimHafez/abdh.txt" 1385 3.227795
cs17 787128350 941071 "http://www.ludvigsen.dhhalden.no/webdoc/levant_servers.html#Lebanon" 0 0.0
cs17 787128351 203370 "http://www.ludvigsen.dhhalden.no/webdoc/ahlan.gif" 0 0.0
cs17 787132565 117453 "http://www.cpsr.org/dox/home.html" 3232 3.803559
cs17 787132569 78375 "http://www.cpsr.org/dox/cpsr.gif" 1644 8.892870
cs17 787132588 528202 "http://www.cpsr.org/dox/memb.form.html" 2644 3.432074
```

Dans cet encadré, Cs17 correspond au nom de l'ordinateur, la série de chiffres 787125748 56906[3] (ligne 1) permet de déterminer exactement le jour, la date et l'heure de connexion liés

[2] Lorsqu'un individu se présente pour la première fois sur un site Internet, un 'mouchard' (cookies) est envoyé via le réseau Internet et copié sur l'ordinateur du visiteur. Ce mouchard correspond à un fichier qui contient entre autres un numéro d'identité personnel. Ainsi, dès que l'ordinateur de l'internaute se connecte à nouveau sur le même site, ce dernier reconnaît qu'il y a eu une ou plusieurs connexions au préalable. Cette technique revient à tenter d'identifier un utilisateur unique, d'une connexion à l'autre.

[3] Temps Unix depuis le 01/01/1970.

à l'URL 'http://www.ludvigsen…', enfin, les informations subséquentes (toujours ligne 1) concernent respectivement le poids du document envoyé[4] (en bytes) et enfin le temps écoulé pour recevoir ce document (en secondes). Puisque le format d'origine de ces données est impropre à une interprétation directe, il est nécessaire de transformer les données de navigation 'brutes' (encadré 1) en informations explicites (encadré 2). C'est à partir de ces informations qu'il est possible de générer des variables explicatives.

Encadré 2 : Restitution des données après traitement de l'information[5]

Individu	Session	Site	Début	Durée	Bytes Echangées
156	1	http://home.mcom.com/	27/01/1995 18:20:01	58 s	147
156	1	http://lycos.cs.cmu.edu/	27/01/1995 18:20:59	2 m 33 s	1 325
156	2	http://www.mit.edu:8001/	21/02/1995 18:11:36	16 m 28 s	40 256
156	3	http://cathouse.org/	21/02/1995 19:12:35	15 m 29 s	6 895
156	4	http://cathouse.org/	22/02/1995 19:45:50	1 m 2 s	1 002
156	4	http://bvp.wdp.com/	22/02/1995 19:46:52	1 h 49 m 43 s	440 296

En supposant qu'une session représente le temps écoulé entre le début et la fin d'une connexion sur Internet (où l'individu est supposé naviguer sur un nombre variable de sites), l'encadré 2 montre par exemple que l'individu 156 a visité le 27/01/1995 deux sites dans sa session numéro un. Il est d'autre part resté 2 minutes et 33 secondes sur Lycos, et a échangé durant cette période un flux d'informations de 1325 Bytes.

Si désormais nous assimilons chaque adresse URL au résultat d'un choix de visite entre différents sites, nous pouvons parler de la navigation sur Internet comme une succession de T_i occasions de choix entre J_i alternatives (les sites Internet), pour un individu *i* donné (ici, les internautes). Plus généralement, nous parlons des comportements de choix *inter-sites* d'un panel de *I* individus, sur une période donnée. Si par contre l'étude s'intéresse aux choix des individus sur un site (les alternatives sont alors les pages visitées ou les produits achetés), nous parlons de comportements de choix *intra-site*.

1.2/ Les données de navigation comme nouvelle source d'information pour l'étude des comportements de choix en ligne.

Les premières études sur Internet se sont concentrée sur la mesure du commerce électronique. Réalisées le plus souvent par des cabinets de conseil, des banques ou des assurances, ces études relèvent un grand nombre d'incohérences dans la mesure du commerce électronique

[4] Pour cette ligne, c'est une image au format GIF qui pèse 1700 bytes.
[5] Fichier non similaire à l'échantillon proposé dans l'encadré 1.

(Brousseau, 2000). Du côté des institutions officielles de statistiques, c'est seulement depuis 1997 que le '*US Census Bureau*' (Newberger, 2001) s'est intéressé aux comportements des ménages ayant un accès Internet à domicile. En France, la première étude portant sur l'état du commerce de détail en ligne a été réalisée en avril 2001 par l'INSEE (Merceron, 2001). Les statistiques fournies restent toutefois agrégées et loin des problématiques relatives aux comportements de choix sur Internet. D'autres indicateurs ont donc été construits par le monde scientifique.

L'analyse des comportements de choix via une interface numérique en réseau a commencé très tôt avec la psychologie cognitive. L'utilisation des données de navigation n'était toutefois pas de mise, et il faudra naturellement se tourner vers le milieu de l'informatique pour voir apparaître une première série d'études. Ce sont deux démarches parallèles, l'une relative à la psychologie cognitive devant des interfaces numériques (Hoffman et al., 1996), et l'autre issue du monde informatique (Glassman, 1994), qui ont inspiré, dès 1997, le physicien (mais aussi économiste) Huberman, à utiliser des données de navigation et observer l'existence d'une loi comportementale générique. Cette *surfing law* stipule que seule une minorité de sites Internet rassemblent une majorité de visites. Reprenant d'autre part un modèle d'options réelles, et s'intéressant par-là aux comportements de visites intra-site […] *there is value in each page that an individual visits, and that clicking on the next page assumes that the information will continue to have some value. Within this formulation an individual will continue to surf until the expected cost of continuing is perceived to be larger than the expected value of the information to be found in the future*[6] (Huberman et al., 1998). Grâce aux données de navigation, les auteurs ont pu déterminer une distribution de probabilité (gaussienne inverse) relative au nombre de pages visualisées sur un site, avant l'arrêt de la navigation. Les observations montrent alors que les individus cessent de naviguer après avoir utilisé en moyenne trois liens (l'écart type est environ de 6), cela quelque soit le site considéré, et pour différents échantillons (de grandeur inégale) à caractéristiques socio-économiques non semblables.

Par la suite, une série d'études croissantes, mais encore restreintes, ont inauguré l'utilisation d'outils plus standards en économie ; retenons par exemple les modèles probabilistes de Fader et al. (2002a)[7], l'utilisation d'un modèle Logit emboîté pour Brynjolfson et al (2000) , et une fusion des deux outils chez Sismeiro et al (2002). Ces trois auteurs se sont concentrés davantage sur les comportements de visite intra-site. Dans ce domaine, d'autres modèles sont

[6] Citation page 1.
[7] Même si Fader a aussi travaillé sur les modèles de choix discrets.

à l'œuvre, retenons par exemple les modèles de comptage (régression négative binomiale, Fader et al., 2002b) relatifs à l'étude des visites répétées sur un site Internet[8].

1.3/ Le mythe de la Nouvelle Economie : marchés 'sans friction' *versus* phénomènes de persistance

Récemment, un résultat important issue de l'étude des comportements de choix entre différents moteurs de recherche, concerne l'observation d'une persistance (Goldfarb, 2002a, 2002b, Drèze et al, 2002).

Heckman (1981) identifie deux sources de persistance (appelées aussi états de dépendance) :

- un état de dépendance purement structurel (*spurious state dependence*) annonçant que la probabilité de choisir une alternative en *t* est majorée lorsque la même alternative a été choisie en *t-1*, dans la mesure où l'utilité s'en trouve majorée ;
- un état de dépendance dû uniquement aux préférences individuelles (*true stata dependence*). En d'autres termes, le choix d'une alternative en *t* ne dépend pas du choix d'une alternative en *t-1*, mais uniquement du fait que d'une occasion de choix à une autre, les individus préfèrent systématiquement la même alternative. Cette alternative n'est donc pas considérée comme un bien d'expérience.

Ce phénomène de persistance des choix (loyauté) est un résultat classique de la littérature marketing relative aux comportements de choix faces à des marques concurrentes (Guadagni et al., 1983). En économie, les explications traditionnelles de tels phénomènes sont solidement ancrées dans l'existence des coûts de changement (Klemperer, 1995). Selon l'auteur, les switching costs *résultent du désir pour un consommateur de rendre compatible son achat courant avec un investissement précédent*[9]. Klemperer assimile donc l'acte d'achat à un investissement qu'il est coûteux de ne pas renouveler pour les raisons suivantes : existence de compatibilité entre produits, construction d'une relation de confiance entre le client et l'entreprise, coût pour apprendre à utiliser un appareil, incertitude sur la qualité d'une autre marque que le consommateur n'a jamais testé, remises ou autres cadeaux consentis aux clients fidèles, et existence de coûts psychologiques pour changer de marque.

Récemment, deux articles de Moshkin et al (2000 a et b) ont proposé une nouvelle source de persistance ; celle concernant l'asymétrie d'information face aux alternatives disponibles. Concrètement, les individus bien informés sur les caractéristiques et le nombre d'alternatives

[8] En France, se sont les études relatives au volume d'audience sur Internet qui ont inauguré une première démarche d'analyse des données de connexions (données Log). Voir pour cela Costes, 2000.

possibles seraient face à des coûts de changement, alors que les individus peu informés auraient des coûts de recherche élevés (Stigler, 1961., Diamond, 1987., Stiglitz, 1989). Même si les auteurs étudient les choix d'un échantillon d'individus entre différents bouquets de chaînes télévisées, il semble que la problématique avancée dans l'article ne soit pas éloignée de celle d'Internet. Alors que la concurrence 'n'est qu'à un seul clic' (les marchés sur Internet étaient supposés sans friction), la littérature observe le plus souvent des addictions fortes aux sites Internet (Johnson et al., 2000).

Ainsi, outre la possibilité d'observer les comportements de choix en ligne, les données de navigation permettent d'implémenter des modèles biens adaptés à l'étude des phénomènes de persistance. Ces derniers appartiennent à l'économétrie qualitative et sont plus largement appelés modèles de choix discrets (Train, 2002., Baltagi, 2001). Une application simple est proposée dans la seconde section.

II. Une application sur le moteur de recherche Yahoo!

2.1/ Analyse descriptive des données de navigation

Les donnés utilisées dans cet article ont été enregistrées dans le cadre d'une étude sur l'optimisation de la vitesse de circulation des informations sur Internet. Le format d'enregistrement, ainsi que le contenu des données se prêtent donc mal à des problématiques économiques. Les données restent toutefois utilisables dans le cadre d'une étude plus restreinte, ici, la persistance des choix. Avec des informations plus riches, il ne sera pas difficile d'extrapoler des modèles de choix plus complexes. L'échantillon initial est composé de 548 étudiants dont 467 ayant navigué en dehors du site de leur université. Nous avons utilisé uniquement les enregistrements pour les mois de janvier et février 1995, les autres mois étant difficilement utilisables. Les caractéristiques de l'échantillon initial sont présentées dans le tableau ci-dessous.

[9] Klemperer, p. 517.

Tableau 1 : Caractéristiques des comportements de navigation

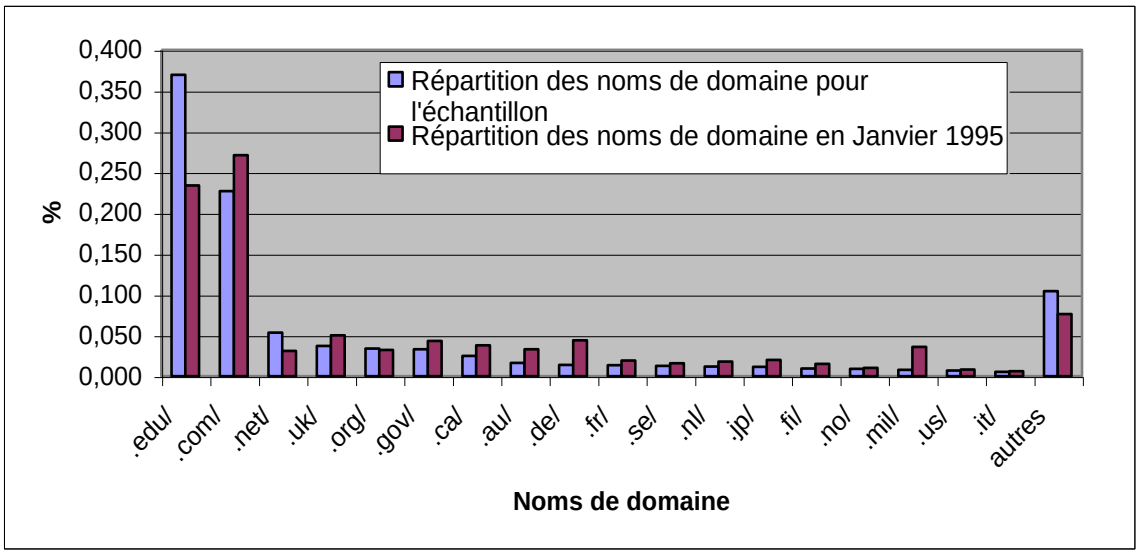
	Nombre de sessions	Nombre de sites visités
Minimum	1	1
Maximum	174	327
Total	5348	15031
Moyenne	11.45	32.18
Ecart type	18.32	43.30

Un écart type supérieur à la moyenne est souvent signe de surdispersion, et par conséquent d'hétérogénéité entre les individus, en terme de nombre de sites visités sur une même période. Une session est définie comme la période entre laquelle un individu se connecte puis se déconnecte pour naviguer sur Internet. Durant chacune des sessions, un individu peut aller visiter un ou plusieurs sites. Notons d'autre part que les sessions se situent à des périodes différentes pour chacun des étudiants. La seule unité de temps commune sont les deux mois d'enregistrements.

Une première approche dans l'analyse descriptive de données traces, consiste à identifier la finalité des navigations. Le fait de connaître précisément l'intitulé des adresses URL nous permet en effet de déterminer le type de navigation. Ainsi, un site enregistré en .edu sera considéré comme relevant de l'éducation américaine. Un site terminé par .net sera davantage tourner vers des activités relatives à Internet, etc. Le nom de domaine .com reste quant à lui relativement générique, étant historiquement l'un des premiers noms de domaine attribués. Nous avons donc estimé que 50 % des sessions ont été utilisées pour naviguer sur le site de l'université. La figure 3 présente la distribution de l'activité de navigation des étudiants[10], en dehors du site de leur université. Cette dernière est comparée avec la distribution des noms de domaine en janvier 1995[11] .

[10] L'annotation 'autres' de la figure 5 désigne les protocoles de communication FTP et News.
[11] Source : http://www.isc.org/ds/WWW-9501/dist-bynum.html

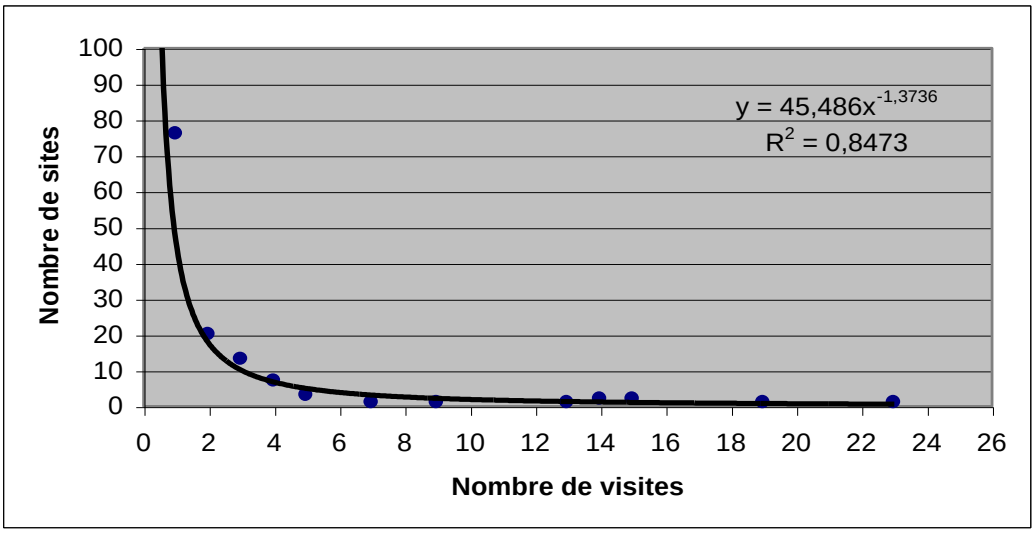
Figure 3 : Répartition des sites visités en fonction des noms de domaine



Alors qu'il y a numériquement moins de sites éducatifs (.edu) que de site estampillés '.com', l'activité première de navigation des étudiants a logiquement un dessein éducatif. Notons d'autre part, qu'en 1995, les outils de recherche étaient majoritairement développés par des pôles universitaires, il y a donc une sur-pondération des .edu. Avec le développement d'Internet, les sites de recherche se sont davantage orientés vers une stratégie marchande, et ont perdu par la même occasion leur .edu.

Une seconde étape peut consister à identifier, et le cas échant différencier, les comportements de navigation individuels. Nous proposons pour cela une variable qui pourrait être un proxy du degré de persistance des individus face à un ensemble de sites Internet. Dans ce sens, notre démarche est identique à celle de Goldfard (2002c). Pour construire ce proxy, nous avons confronté le nombre de visites au nombre de sites (figure 4).

Figure 4 : Distribution du nombre de visites *versus* le nombre de sites pour l'individu 160



Si nous regardons les deux points extrêmes de cette figure, nous voyons pour le quart nord-ouest, que l'individu *i* = 160, a visité chacun des 76 sites une seul fois, [point de coordonnée (1, 76)], avec X, le nombre de visites et Y, le nombre de sites. A l'extrémité du nuage de points (quart sud-est), l'individu 160 a visité 23 fois un même site (ici http://www.timeinc.com/), point de coordonnée (23, 1). Dès lors, un nombre important de points dans le quart sud-est, relativement au quart nord-ouest et sud-ouest, peut être assimilé à un comportement de navigation peu volatil pour l'individu 160. En d'autres termes, ce dernier a tendance à revenir régulièrement sur certains sites. Il semble que le coefficient *a* de la fonction puissance ajustée $Y = cX^{-a}$ (ou encore la pente *a* de la régression Log-Log) puisse être un bon proxy du degré de volatilité individuel[12]. Pour vérifier cela, nous avons en effet calculé les coefficients *a* pour chaque individu, et comparé ces coefficients avec la distribution respective du nuage de points (plus ou moins semblable à la figure 6) traduisant les comportements de navigation individuels. Une pente faible [notamment proche de (-1) avec un écart type de 0,5] montre clairement que les individus ont été peu volatils sur la période d'étude considérée. Une pente qui augmente procède à la logique inverse.

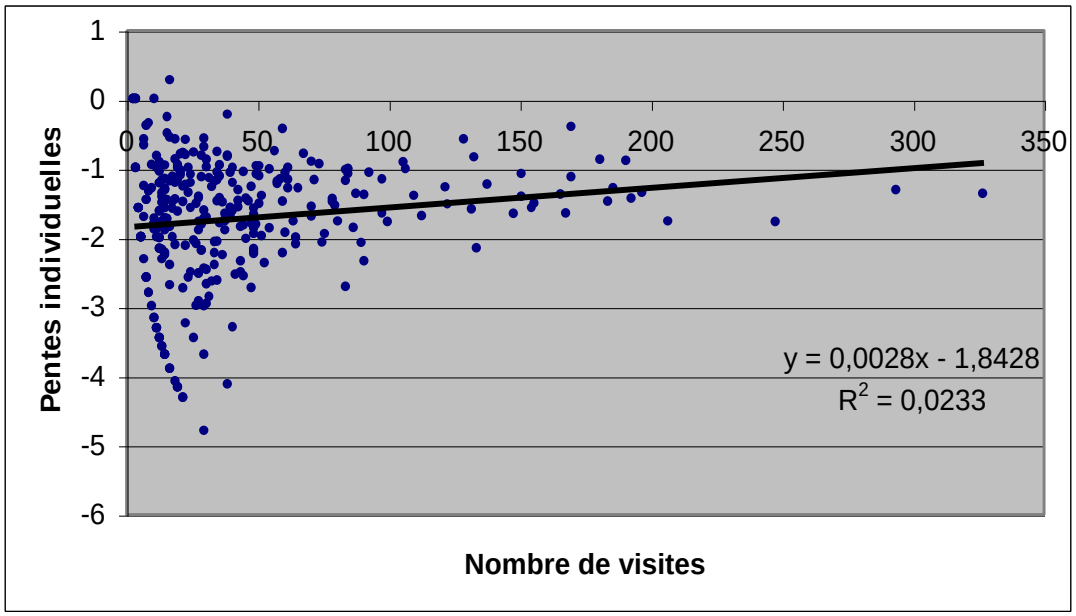
La figure 5 trace le nuage des pentes pour les 332 individus[13], relativement au nombre de visites effectuées. Nous observons une faible dépendance linéaire entre le nombre de visites et

[12] L'utilisation des fonctions et lois puissances n'est pas nouveau en économie, cf. **Gibrat, R**., (1931), 'Les inégalités économiques', Recueil Sirey, Paris.

[13] Si un individu a navigué une fois sur chaque site, ou visité *n* fois un même site durant la période de l'étude, la pente ne peut être calculée. Nous n'avons donc pas pris en compte ces individus. Cela avait peu d'incidence sur nos résultats.

les pentes respectives de chaque individu. Malgré tout, il semble que la variabilité des pentes diminue avec le nombre de visites.

Figure 5 : Spectre comportemental de navigation en fonction des visites



Notons enfin que la pente moyenne est de -1,71 (écart type égal à 0,87) et que le coefficient de détermination moyen des ajustements est de 0,86 avec un écart type de 0.16.

2.2/ Un modèle Logit sur données de panel : test de la persistance sur le site Yahoo!

Nous avons utilisé dans notre échantillon les 332 individus appartenant au 'spectre comportemental' de la figure 5. L'objectif est d'implémenter un modèle Logit multinomial sur données de panel, afin de tester l'existence ou non d'une persistance sur le moteur de recherche Yahoo! Notons que les études actuelles utilisent largement les moteurs de recherche, puisque ce type de sites est le plus souvent incontournable lorsque l'on navigue sur Internet, et rassemble dans ce sens un grand nombre de connexions, donc de données de navigation. Avec le développement d'Internet, la communauté scientifique disposera à l'avenir de suffisamment de données pour effectuer des recherches sur les comportements de choix des internautes faces à des sites de commerce électronique concurrents.

Le modèle Logit est largement utilisé dans l'analyse des données de navigation. Il y a selon nous plusieurs raisons à cela. Le modèle Logit peut être considéré comme l'outil de base en économétrie des choix discrets. S'il contient des hypothèses restrictives, notamment

l'hypothèse IIA (independence of irrelevant alternatives), ces dernières peuvent rapidement être relâchées pour tendre vers des spécifications sans doute plus réalistes. Comme l'étude des comportements en ligne reste relativement nouvelle en économie, une première démarche de 'dégrossissage' peut justifier l'utilisation d'un modèle générique. Une autre raison peut être historique. La littérature marketing a longuement travaillé sur ce types de modèles pour analyser les données dites 'scanner'. Ces dernières sont issues des supermarchés et contiennent les enregistrements systématiques des achats pour un panel de clients. Les données scanner étant voisines des données de navigation, la transposition des outils d'analyse a Internet est devenu naturelle. Enfin, la raison sûrement la plus importante réside dans l'extrême richesse (potentielle) des données de panel, lorsque celle-ci concentrent à la fois les données de navigation, les caractéristiques socio-économiques des internautes, mais aussi les caractéristiques des alternatives elles-mêmes. Ces dernières sont d'autant plus faciles à enregistrer qu'elles sont consultables sur Internet. Il n'est en effet pas difficile aujourd'hui de relever quotidiennement les prix d'un produit sur différents sites. Un tel volume d'informations, couplées à une logique dynamique des choix (dans la meilleure situation, il est possible de prendre en compte l'histoire entière des choix), augmente de façon considérable le nombre d'observations. Dès lors, pour minimiser le temps de calcul des logiciels de statistiques, il est préférable d'adopter une forme analytique simple pour un modèle de choix discret, ce qui est le cas du Logit.

Le modèle général est spécifié comme suit : à chaque occasion de choix *t*, un individu *i* sélectionne l'alternative *j* qui lui offrira la plus grande utilité U_{ijt}. Il y a au total *I* individus, J = 2 alternatives et T_i occasions de choix. Le modèle est donc un Logit binomial. Le caractère panel des données suppose l'existence d'un effet individuel α_i . Par défaut, nous suposons une distribution aléatoire des effets individuels: $\alpha_i \sim IID(0, \sigma_\alpha^2)$. Le modèle est donc à effets aléatoires.

Tout au long de sa navigation, et pour chaque occasion de choix, l'individu *i* a deux alternatives : visiter le moteur de recherche Yahoo! ou visiter un autre site. Dès lors, la probabilité de choisir le moteur de recherche Yahoo! est définie par :

$$P(y_{it} = 1) = \frac{e^{\alpha_i + X_{it}'\beta}}{1 + e^{\alpha_i + X_{it}'\beta}}$$

Nous intégrons dans la partie observée de l'utilité X_{ijt}, quatre variables explicatives :

- la variable *session* qui enregistre le numéro de session de l'individu *i* à l'occasion de choix *t* ;
- la variable *TempsSession* qui enregistre le temps passé sur la session en cours ;
- la variable *indice* qui est la pente *a* calculée pour construire la figure précédente ;
- la variable *TempsConex* : qui est le temps passé sur le site considéré.

Pour simplifier, nous supposons que la variable temps est exogène au modèle.

Les résultats de l'estimation pour la totalité des variables explicatives sont rendues dans le tableau ci-dessous.

```
. xtlogit site session TempsSession indice TempsConex, i(indiv) re nolog

Random-effects logit                            Number of obs      =     17929
Group variable (i) : indiv                      Number of groups   =       332

Random effects u_i ~ Gaussian                   Obs per group: min =         6
                                                               avg =      54.0
                                                               max =       360

                                                Wald chi2(4)       =     97.83
Log likelihood  = -3418.4578                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
        site |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     session |   .0010664   .0025676     0.42   0.678    -.0039661    .0060989
TempsSession |  -.0001622   .0000283    -5.73   0.000    -.0002177   -.0001067
      indice |  -.3045704   .1162943    -2.62   0.009    -.5325031   -.0766377
  TempsConex |   .0007397   .0000811     9.12   0.000     .0005807    .0008988
       _cons |  -2.722675   .2170545   -12.54   0.000    -3.148094   -2.297256
-------------+----------------------------------------------------------------
    /lnsig2u |   .0105701   .1507193                     -.2848343    .3059746
-------------+----------------------------------------------------------------
     sigma_u |   1.005299    .075759                      .8672594     1.16531
         rho |   .2350023   .0082361                      .1860805     .292169
------------------------------------------------------------------------------

Likelihood ratio test of rho=0: chibar2(01) =   303.64 Prob >= chibar2 = 0.00
```

Tous les coefficients ont le signe attendu. Une augmentation du nombre de sessions procède à une hausse de la probabilité d'utiliser à nouveau le site Yahoo!. Cela est le cas lorsqu'il y a persistance des choix. L'intervalle de confiance reporte toutefois un signe négatif, l'écart type est élevé et la probabilité Z n'est pas à l'avantage de la variable *session*. Il semble que l'erreur vienne du traitement même de cette variable qui ne dépend pas suffisamment des occasions de choix. L'utilisation du nombre total de sessions aurait semble-t-il été plus judicieux. Pour la seconde variable, une augmentation du temps sur une session provoque une diminution de la probabilité de visiter le moteur de recherche. Cela n'est pas étonnant, puisqu'un moteur de recherche est un site où l'on ne reste que provisoirement, son premier objet étant d'offrir des adresses pertinentes à l'internaute. Le signe négatif de la variable *indice* correspond à nos observations précédentes : un coefficient proche de la valeur 1 stipulait un comportement général d'addiction élevé envers les sites Internet. Dès lors, une augmentation de ce coefficient signifie une diminution de la probabilité d'utiliser à nouveau le moteur de recherche Yahoo! . Enfin, plus le temps passé sur le site Yahoo! est élevé, et plus la

probabilité d'y revenir est forte. La loyauté semble donc être une variable explicative importante dans la prévision des futures usages d'un individu sur Internet. L'utilisation d'une variable de retard $y_{i,t-1}$ serait à tester dans d'autres modèles.

Les estimations suivantes se déchargent de la variable *session*, puis de la variable *TempsSession*. Les coefficients restent stables et du signe attendu.

```
Log likelihood  = -3418.5402                    Prob > chi2        =    0.0000
------------------------------------------------------------------------------
        site |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
TempsSession |  -.0001631   .0000282    -5.78   0.000    -.0002184   -.0001078
      indice |  -.3030735   .1172529    -2.58   0.010     -.532885   -.0732621
  TempsConex |   .0007405   .0000812     9.12   0.000     .0005814    .0008995
       _cons |  -2.710606   .2190014   -12.38   0.000    -3.139841   -2.281371
-------------+----------------------------------------------------------------
    /lnsig2u |  -.0005047   .1473357                     -.2892774    .2882679
-------------+----------------------------------------------------------------
     sigma_u |   .9997477   .0736493                      .8653349    1.155039
         rho |   .2330172   .0080039                      .1854086    .2885207
------------------------------------------------------------------------------
Likelihood ratio test of rho=0: chibar2(01) =   337.30 Prob >= chibar2 = 0.000
```

```
Log likelihood  = -3437.8605                    Prob > chi2        =    0.0000
------------------------------------------------------------------------------
        site |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      indice |  -.3583667   .1093988    -3.28   0.001    -.5727844    -.143949
  TempsConex |   .0005488   .0000731     7.51   0.000     .0004055    .0006921
       _cons |  -2.768242   .2042391   -13.55   0.000    -3.168543   -2.367941
-------------+----------------------------------------------------------------
    /lnsig2u |  -.0355615   .1417195                     -.3133266    .2422036
-------------+----------------------------------------------------------------
     sigma_u |   .9823764   .0696109                      .8549919     1.12874
         rho |   .2268106   .0075544                      .1818038    .2791576
------------------------------------------------------------------------------
Likelihood ratio test of rho=0: chibar2(01) =   339.88 Prob >= chibar2 = 0.000
```

Conclusion : futures voies de recherches

L'article s'est chargé de promouvoir l'utilisation des données de navigation pour analyser les comportements des internautes. Un modèle Logit a été proposé comme outil d'analyse générique, mais aussi afin de relever un fait paradoxale mais semble-t-il coutumier, relatif à la persistance potentielle d'une majorité d'internautes sur certains sites Internet.

Le modèle Logit proposé reste toutefois très insuffisant. Pour améliorer ce dernier dans sa structure, il sera nécessaire de considérer un nombre plus large d'alternatives (Logit Multinomial), tout en s'affranchissant de l'hypothèse IIA implicite au modèle (Logit Multinomial emboîté). D'autre part, une ou plusieurs variables de retard devraient être intégrées.

Du côté des données utilisées, l'étude des choix serait grandement améliorées si nous disposions d'informations aussi bien sur les internautes que sur les sites.

Bien d'autres améliorations pourront être envisagées, d'autant plus que récemment, un véritable programme de recherche sur l'analyse des comportements de choix en ligne (utilisant des données de navigation) a été initié aux Etats-Unis (Bucklin et al., 2002). Les auteurs proposent ainsi le développement d'outils d'analyses autres que les modèles de choix discrets.

Bibliographie :

Baltagi, B. H., (2001), 'Econometric analysis of panel data', John Wiley & Sons, 293 p.

Brousseau, E., (2000), 'Commerce électronique : ce que disent les chiffres et ce qu'il faudrait savoir', *Economie et Statistique*, n° 339-340, 9/10, pp. 147-170.

Brynjolfsson, E., Smith, M, D., (2000), 'The great equalizer ? The role of shopbots in electronic markets', Working Paper, Sloan School, MIT.

Bucklin, R.E., Lattin, J. M., Ansari, A., Bell, D., Coupey, E., Gupta, S., Little, J. D. C., Mela, C., Montgomery, A., Steckel, J., (2002), 'Choice and the Internet : from clickstream to research stream', U. C. Berkeley 5th Invitational Choice Symposium.

Costes, Y., (2000), 'Comprendre et mesurer le profil et le comportement des internautes', *Revue Française du Marketing*, n° 177/178, 2-3, pp. 153-167.

De Palma, A., Thisse, J. F., (1989), 'Les modèles de choix discrets', Working Paper, BETA, 89/01.

Diamond, P., (1987), 'Search Theory', in Eatwell, J., Milgrate, M., Newman, P., (Eds)., *The New Palgrave*, 4, London : Macmillan, pp. 273-279.

Drèze, X., Lee, S., Zufryden, F., (2002), 'A study of consumer switching behavior across Internet portal websites', *Under Revision at the International Journal of Electronic Commerce*.

Fader, P. S., Moe, W., (2002a), 'Capturing Evolving Visit Behavior in Clickstream Data', Under first review at *Management Science*.

Fader, P., Bradlow, E. T., Hardie, B. G. S., (2002b), 'Bayesian Inference for the Negative Binomial Distribution via Polynomial Expansions', Forthcoming, *Journal of Computational and Graphical Statistics.*

Goldfarb, A., (2002a), 'State dependence at internet portal', Joseph L. Rotman School of Management, University of Toronto, working paper.

Goldfarb, A., (2002b), 'Analysing website choice using clickstream data', Joseph L. Rotman School of Management, University of Toronto, working paper.

Goldfarb, A., (2002c), 'Interpreting Internet clickstream data', Joseph L. Rotman School of Management, University of Toronto, working paper.

Guadagni, P. M., Little, D. C., (1983), 'A logit model of brand choice calibrated on scanner data', in *Marketing Sciences*, 2(3).

Heckman, J. J., (1981), 'Statistical models for discrete panel data', in Manski & McFadden Eds, Structural analysis of discrete data with econometric applications, 477 p.

Hoffman, D. L., Novak, T. P., (1996), 'Marketing in hypermedia computer-mediated environments : conceptual fondations' in Journal of Marketing, Vol 60, Issue 3.

Huberman, B. A., Lukose, R. M., (1998), 'Surfing as a real option', ICE 98, ACM Press.

Huberman, B. A., Pirolli, P. L. T., Pitkow, J. E., Lukose, R. M., (1997), 'Strong regularities in the World Wide Web surfing', *Science*, April 3, Vol 280, n° 5360, pp. 95-97.

Johnson, E. J., Lohse, J., Bellman, S., (2000), 'Cognitive lock-in', Columbia University Business School. Mimeographed.

Klemperer, P., (1995), 'Competition when Consumers have Switching Costs', *Review of Economic Studies*, 62, October.

Lynch, J. G., Ariely, D., (2000), 'Wine online : search costs affect competition on price, quality, and distribution', in Marketing Science, Vol 19.

Merceron, S., (2001), 'Le commerce de détail s'initie à la vente sur Internet', INSEE Première, n° 771, Avril, 4 p.

Moshkin, N., Shachar, R., (2000a), 'Switching Cost or Search Cost?' ,The Foerder Institute for Economic Research Working Paper No. 3-2000, January.

Moshkin, N., Shachar, R., (2000b), 'The Asymmetric Information Model of State Dependence', Working Paper, Yale University.

Newburger, E. C., (2001), 'Home computers and Internet use in the United States', US Census Bureau, Current Population Reports, US Department of Commerce, September issue, 9 p.

Pénard, T., (2002), 'Mythes et réalités du commerce électronique : une revue des études empiriques', Baslé, M., Pénard, T. Eds., 'eEurope. La société européenne de l'information en 2010'.

Sismeiro, C., Bucklin E., (2002), 'Modeling Purchase Behavior at an E-Commerce Web Site: a conditional probability approach', *Working Pape*r, Anderson School, UCLA.

Smith, M. D., Bailey, J., Brynjolfsson, E., (1999), 'Understanding digital markets : review and assessment'

Stigler, G. J., (1961), 'The economics of information', *Journal of Political Economy*, Vol. LXIX, n° 3, June, pp. 213-225.

Stiglitz, J. E., (1989), 'Imperfect information in the product market', Handbook of Industrial Organization, Volume I, Schmalensee, R. & Willig, R. D. (Eds), pp. 771-847.

Train, K., (2002), 'Discrets choice methods with simulation', Cambridge University Press.