UNIVERSITE DE RENNES 1

Le Saux Loïc
Marot Gildas
Tanguy Brewal

Etude sur les medias en France

Analyse de données – M1 ISC – Mars 2008

Sommaire

Introduction

Nous vivons dans un monde où la présence des médias est importante. Que ce soit par la télévision, le papier, la radio, Internet et les Technologies de l'Information et de la Communication, nous en utilisons au moins un par jour.

Il nous a alors semblé pertinent dans le cadre de ce mémoire d'analyse de données d'étudier différents aspects de ces médias.

Notre analyse va donc porter sur la presse écrite, avec les habitudes des lecteurs et sur Internet, sur différents sites afin de savoir s'ils correspondent à leur cible, ainsi que sur les entreprises de services utilisant les diverses fonctionnalités offertes par Internet.

Dans un premier temps, nous présenterons les méthodes qui seront utilisées dans cette analyse, avant de passer à l'analyse proprement dite.

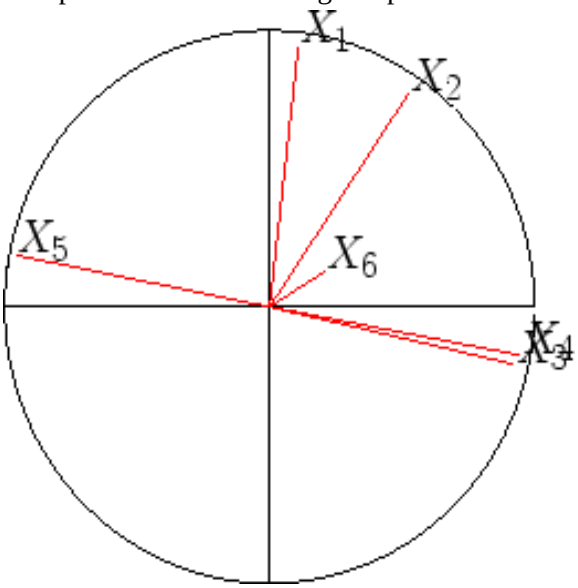
Présentation de l'ACP

Généralités

L'analyse en composantes principales, notée ACP, est utilisée pour synthétiser une volumineuse information contenue dans un tableau de mesure. Cette analyse permet de déterminer s'il y a des liens entre les variables étudiées. En effet synthétiser l'information permet de découvrir les principales structures internes à ces données.

Principes de l'ACP

Le principe de l'ACP est de représenter les individus par des points de l'espace vectoriel R^p. Ces points forment un nuage de points.



L'ACP représente aussi les variables par des vecteurs au sein d'un cercle de corrélation.

Ce cercle de corrélation permet de juger de l'inertie de chaque variables dans l'analyse, en effet, plus le vecteur est proche du bord du cercle, plus la quantité d'information restituée est élevée. X_1 a plutôt une représentation parfaite alors que X_6 est peu restituée sur les axes sélectionnés.
Nous pouvons aussi déterminer les corrélations entre chaque variable.
Si elles ont la même direction et le même sens, elles sont fortement corrélées positivement (X_1 et X_3).
Si au contraire elles ont un sens opposées, alors elles sont corrélées négativement (X_5 et X_3).

Le cas où les variables ne sont pas corrélées est lorsque les vecteurs forment un angle droit (X_1 et X_5).

L'ensemble est ensuite représenté dans un graphique regroupant le cercle de corrélation et le nuage de points des individus : le biplot.

L'analyse par l'ACP par la méthode ABC

La règle A : On mesure la qualité globale de l'ajustement de chaque nuage, par le pourcentage d'inertie obtenue en projection sur les plans représentés. Un minimum de 60 % de l'information doit être contenu par l'ensemble des axes. Mais, pour des soucis de représentation, un minimum de deux axes principaux sera utilisé : F1 et F2, ce pour des raisons évidente de représentation dans l'espace, via un plan suffisamment logique et lisible. Ainsi, d'autres axes pourraient être utiles afin de renforcer les interprétations mais, en règle générale, pour des raisons dimensionnelles, une représentation de plus de

deux plans n'est pas pratique. On peut même descendre jusqu'à 40% d'inertie dans le cas de gros tableaux mais les graphiques indiquent alors une tendance et non des faits certains.

La Règle B : On étudie le cercle de corrélation des variables, et on montre que les points du nuage des variables se trouvent à l'intérieur du cercle. On étudie alors les corrélations comme vu précédemment et l'on peut regrouper des paquets de variables très corrélées par une nouvelle variable appelée composante principale.

La Règle C : Il s'agit ici d'interpréter les axes et c'est précisément cette interprétation qui révèle les structures internes aux données. Elle se fait à l'aide du biplot. Géographiquement le milieu représente la moyenne et l'on peut alors projeter les individus sur les axes sélectionnés pour établir une analyse et créer des groupes d'individus, nous révélant des tendances.

Il faut ensuite relire les données du tableau de mesure initial à la lumière des résultats de l'ACP et vérifier si le tout est cohérent.

Présentation de l'Analyse Factorielle des correspondances

Généralités

L'analyse factorielle des correspondances, appelé AFC, est une méthode statistique d'analyse des données visant à rassembler en un nombre réduit de dimensions la plus grande partie de l'information initiale en s'attachant non pas aux valeurs absolues mais aux correspondances entre les variables, c'est-à-dire aux valeurs relatives. Cette réduction est d'autant plus utile que le nombre de dimensions initial est élevé. L'AFC offre la particularité (contrairement aux ACP) de fournir un espace de représentation commun aux variables et aux individus. Pour cela l'AFC raisonne à partir de tableaux réduits ou de fréquences.

On utilise l'AFC pour synthétiser diverses informations, particulièrement volumineuse dans un tableau de contingence. Ce dernier correspond à un tableau à double entrée, les caractères observés sur une même population sont donc représentés simultanément.

Pour une population de n individus, c'est un tableau qui contient dans la case située en ligne i et en colonne j, l'effectif n_{ij} d'individus qui présentent la modalité i d'un premier caractère noté I, et la modalité j d'un second caractère J. L'AFC sert donc à bien déterminer et à hiérarchiser toutes les dépendances entre les lignes et les colonnes du tableau. La synthèse de l'information consiste essentiellement à rendre compte des liaisons existant entre les caractères I et J. Si les distributions conditionnelles ne sont pas différentes des distributions marginales, alors ces deux caractères sont dépendants. On indique donc les modalités originales et leur lien aux modalités de l'autre caractère.

Principes de L'AFC

Le principe de la méthode est de partir sans à priori sur les données et de les décrire en analysant la hiérarchisation de l'information présente dans les données. L'analyse factorielle des correspondances AFC développée emploie la métrique du chi-deux. C'est une ACP faite avec une métrique particulière sur le tableau des profils lignes (on y retrouve les fréquences de la distribution conditionnelle selon le caractère J, sachant la modalité i observée pour le caractère I). Il existe donc

quelques différences. Ici, chaque ligne est affectée d'une masse qui est sa somme marginale, le tableau étudié est le tableau des profils des lignes, ce qui permet de représenter dans le même espace à la fois les deux nuages de points associés aux lignes et aux colonnes du tableau de données. La représentation est quasi-similaire à celle d'une ACP des profils colonnes (On y retrouve les fréquences de la distribution conditionnelle selon le caractère I sachant la modalité j observée pour le caractère J). On retiendra pour l'AFC une représentation moyenne de celles que donnent, d'une part les ACP des profils lignes, d'autre part les ACP des profils colonnes.

Objectifs de L'AFC

Les objectifs de l'AFC sont multiples. Les principaux objectifs de cette analyse sont de connaitre l'organisation des données, et d'en connaitre la configuration après analyse. Un des objectifs est de produire une représentation, dans un repère unique, des catégories en lignes et en colonnes du tableau afin de mettre en évidence leurs positions respectives, et les éventuelles attractions, répulsions entre les caractéristiques.

Lieux et Horaires de Consultation de la Presse Écrite en France

A) *Présentation des deux tableaux « Presse Écrite »*

1) *les tableaux*

Consultation de la presse écrite au cours de la journée (*sur 100 lecteurs de la veille)*

	avant 8h	entre 8h et 10h	entre 10h et midi	De midi à 14 heures	entre 14h et 18h	après 18h
quotidiens régionaux	17	28	23	22	24	30
quotidiens nationaux	11	25	24	24	29	40
quotidiens urbains gratuits	24	35	14	19	21	26
hebdo regionnaux	4	13	20	18	30	36

Lieux de consultation de la presse écrite (*sur 100 lecteurs de la veille)*

	Domicile	Transports	Lieu de travail	Parents, Amis
quotidiens régionaux	71	1	13	8
quotidiens nationaux	65	5	19	3
quotidiens urbains gratuits	27	43	30	1
hebdo regionnaux	76	1	7	10

2) *Source*

Ces tableaux sont le résultat d'une enquête réalisé par TNS Sofres pour le compte de l'EPIQ (Étude de la Presse d'Information Quotidienne). Ce sont deux organismes reconnu, on peut donc considérer les données comme fiable. 25484 interviews ont été réalisées de janvier à décembre 2007.

3) *Les Individus*

Les individus de ces tableaux sont différents types de journaux. Les quotidiens régionaux comme par exemple Ouest-France, Sud-Ouest, La Voix du Nord. Les quotidiens nationaux comme La Croix, Les Echos, L'Equipe. Les quotidiens urbains gratuits comme Metro ou 20 Minutes. Et enfin les hebdos régionaux comme L'Equipe Dimanche, Le Journal du Dimanche, Aujourd'hui en France Dimanche…

4) *Les Variables*

Les variables du premier tableau sont les heures de lecture au cours de la journée qui est découpé en 6 périodes. Les variables du second tableau sont les lieux de lecture des différents journaux.

5) *La Problématique*

Il est intéressant d'analyser ces deux tableaux et de les mettre en parallèle afin de mieux comprendre les habitudes de lecture de la presse en France.

B) Choix de la méthode

Chaque méthode étudiée en analyse de données correspond à un type de tableau. En effet le tableau de mesure est associé à l'ACP et le tableau de contingence à l'AFC, par conséquent nous allons donc recourir ici à l'ACP. Les principes et objectifs de cette méthode ont été présentés dans la partie 1 de ce dossier.

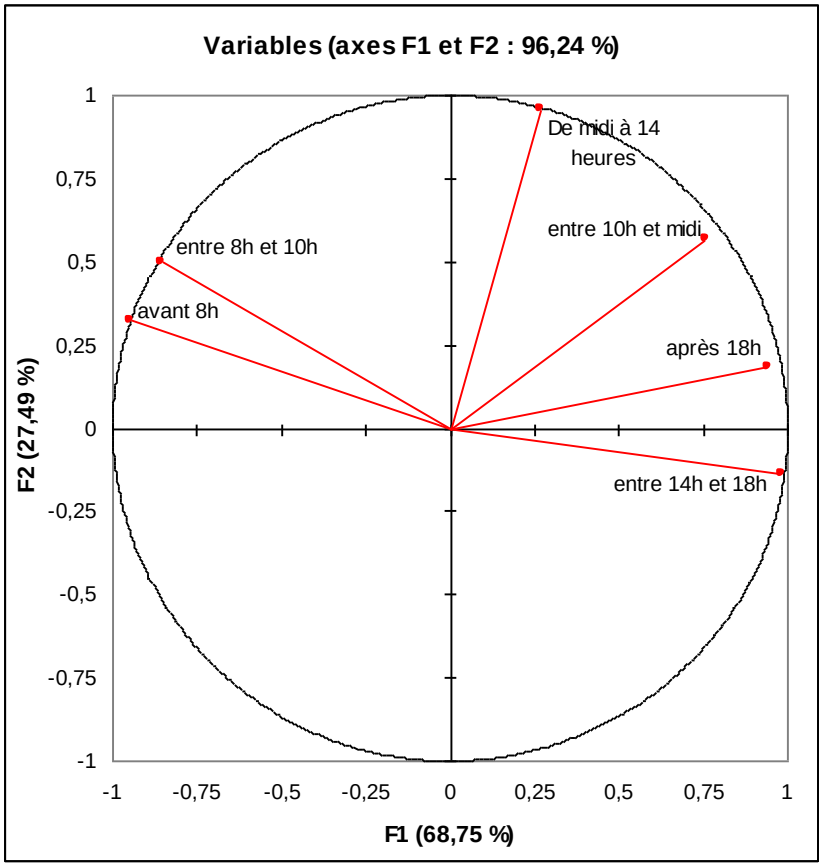
C) Analyse de l'ACP via XLstat

a) Tableau « Consultation de la presse écrite au cours de la journée »

Règle A

96,24% de l'information est réuni sur les axes F1 (68,75%) et F2 (27,49%). On peut observer dans les annexes la contribution de chaque axe à l'inertie, et l'on remarque qu'il y a un décroché entre l'axe F2 et l'axe F3. C'est pourquoi il n'est pas intéressant de retenir d'autres axes que les axes F1 et F2.

Règle B



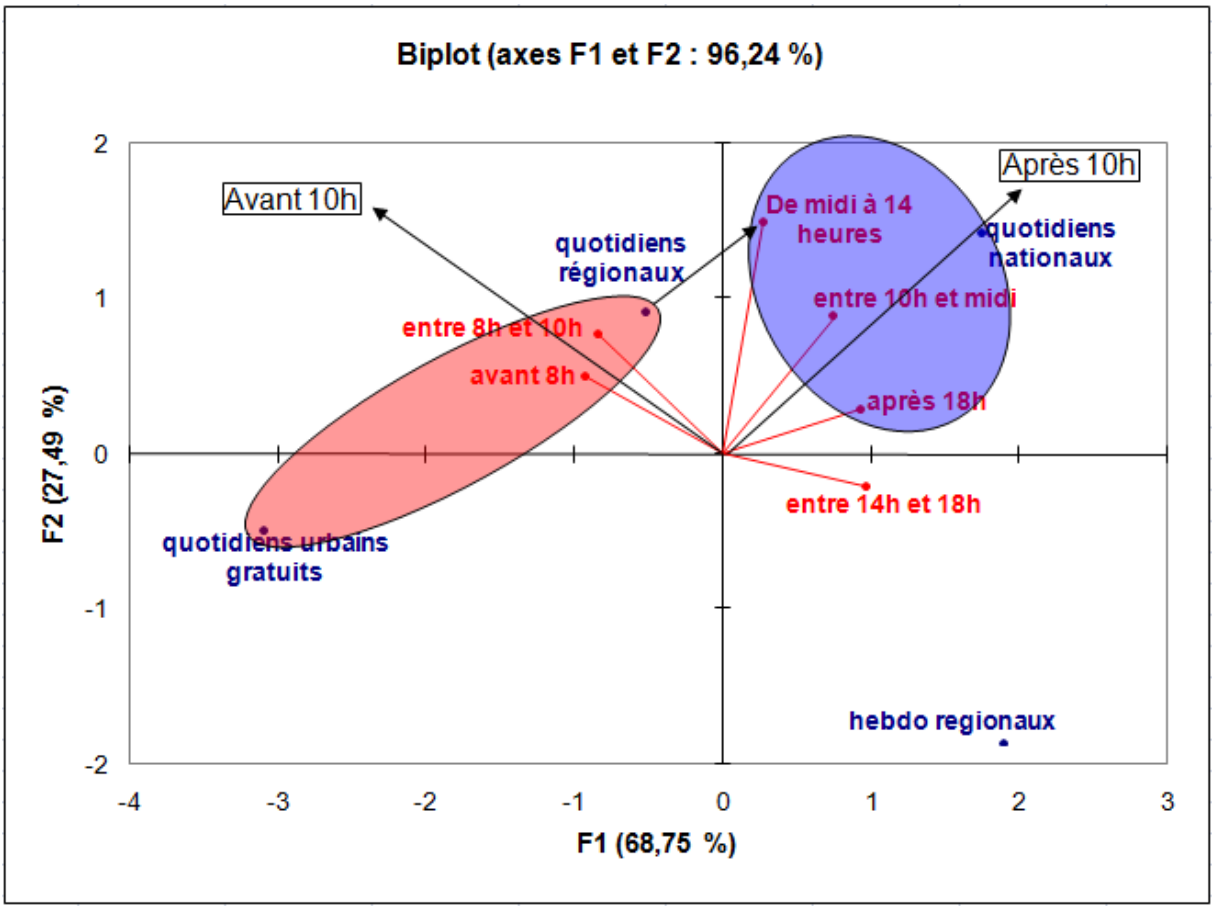
Le cercle des corrélations sert à définir les vecteurs des composantes des individus sur les axes principaux d'inertie. Les points du nuage des variables se trouvent importés à l'intérieur du cercle des corrélations. La corrélation est d'autant plus forte que les variables vont dans le sens de l'axe. Chaque variable est corrélée selon la part de chaque période par rapport aux autres.

Plus les variables sont situées au bord du cercle, mieux elles sont représentées. Les corrélations des variables avec les axes favorisent la formation de composantes principales. Une composante principale est un *« paquet »* de variables très corrélées.

Dans le cadre de notre analyse, la plupart des vecteurs joignent le cercle. Nous pourrons donc obtenir des conclusions parfaites. Plus les coordonnées des variables coïncident avec le cercle (se rapprochent de 1), plus les interprétations sont pertinentes.

On peut déjà séparer les individus en deux groupes sur l'axe F1, à gauche la lecture tôt le matin et à droite la lecture l'après midi et le soir. On remarque aussi que les vecteurs « avant 8h » et « de midi à 14h » sont perpendiculaires donc qu'il n'y a pas de corrélations entre eux.

Règle C



On peut ici dégager un axe de « lecture avant 10 heures » et un axe de « lecture après 10 heures ». On observe que les quotidiens urbains gratuits sont surtout lu le matin c'est-à-dire au moment de leur distribution. Cependant Direct Soir distribué en fin de journée n'a pas été inclus dans ces enquêtes, il sera inclus dans l'enquête pour 2008 ce qui devrait changer la donne.

On observe aussi que les quotidiens régionaux sont lus le matin mais aussi de midi à 14 heures. Donc sans doute sur les temps de pause durant la journée c'est-à-dire avant de partir au travail et pendant la pause déjeuné.

Les quotidiens nationaux quant à eux sont lus de 10 heures à 14 heures et après 18 heures, donc plus tardivement dans la journée.

Il serait sans doute plus pertinent de situer la lecture des hebdos régionaux au cours de la semaine plutôt que sur une journée. Cependant on observe qu'ils sont plus consultés entre 14h et 18h.

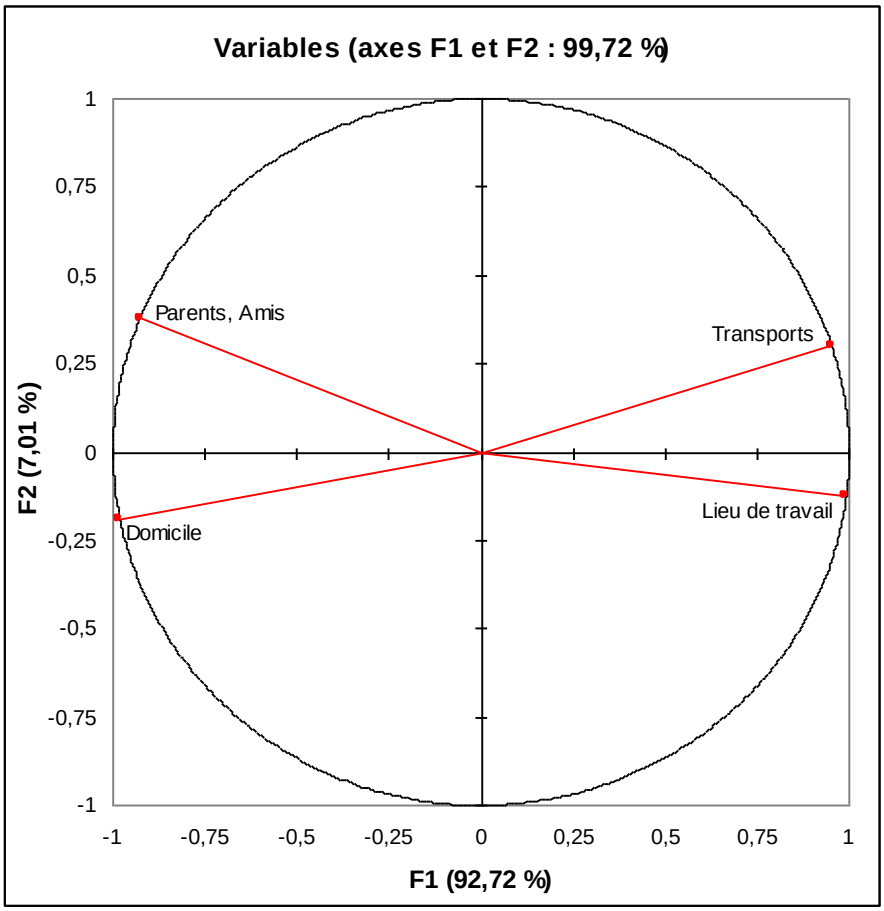
On notera que la contribution à l'axe F2 des variables de 8 heures à 14 heures est de plus de 90%. Le début de mâtiné et l'après midi sont donc faiblement expliqué par cet axe.

b) tableau « Lieux de consultation de la presse écrite »

Règle A

99,72% de l'information est réuni sur les axes F1 (92,72%) et F2 (7,01%). On peut observer dans les annexes la contribution de chaque axe à l'inertie, et l'on remarque qu'il y a un décroché entre l'axe F2 et l'axe F3. C'est pourquoi il n'est pas intéressant de retenir d'autres axes que les axes F1 et F2.

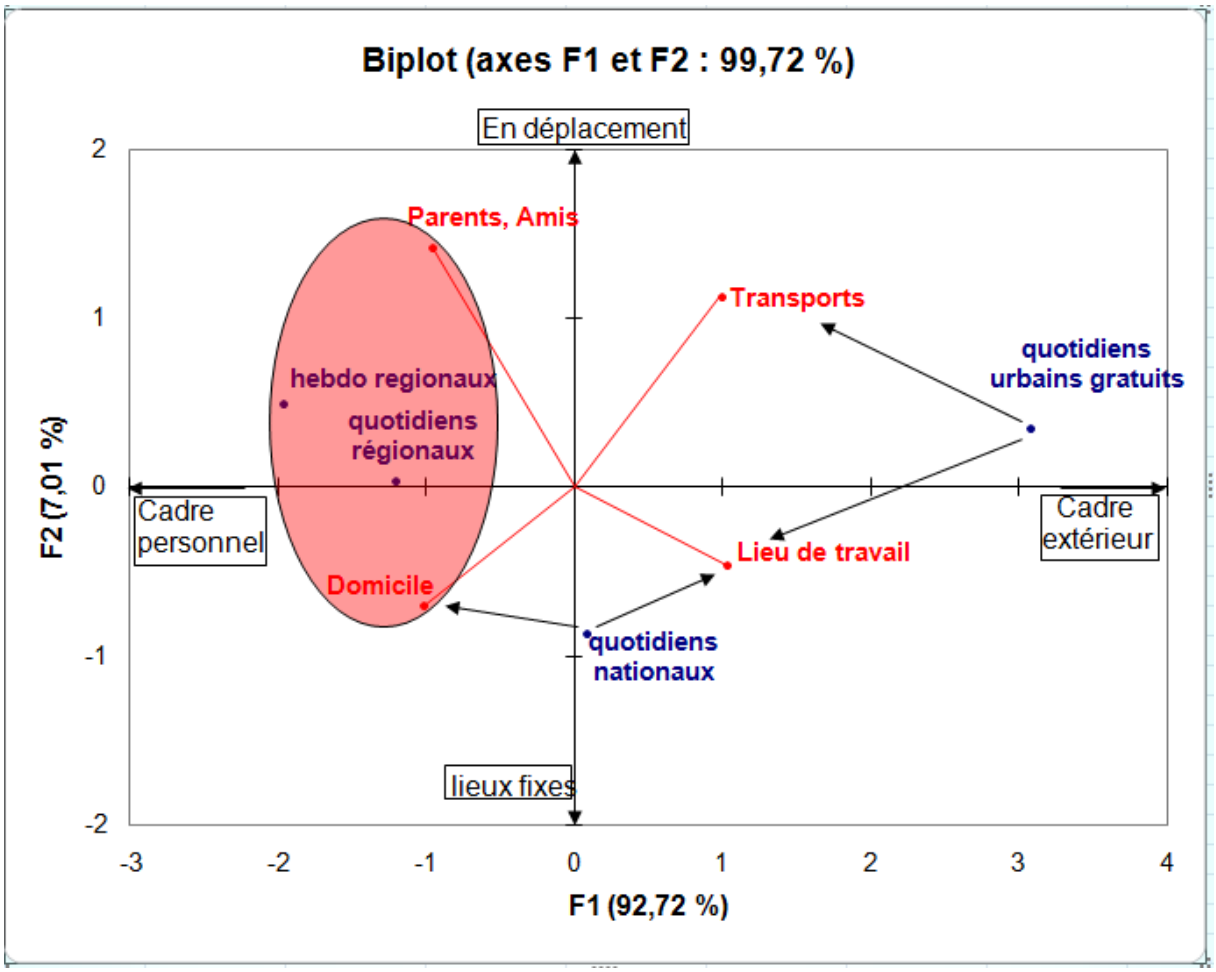
Règle B



On remarque ici que « parents, amis » est corrélé négativement avec « lieu de travail » et que « domicile » est corrélé négativement avec « transport ». On remarque que la bissectrice des deux premiers vecteurs et des deux suivants correspond à l'axe F1 qu'on pourrait alors interpréter comme un axe « cadre personnel/ cadre extérieur ».

On peut séparer les individus en deux groupes sur l'axe F2. En bas nous avons le domicile et le lieu de travail qui sont des endroits fixes ou les personnes peuvent lire de manière régulière et à priori tranquillement (tout dépend du travail). En haut nous avons des lieux de lecture plus instable. A part en région parisienne un trajet dure rarement très longtemps, on peut être amené à lire debout et on peut être coupé au cours de sa lecture. De même chez les parents et amis on lit à priori un quotidien qu'on n'a pas choisi et on peut facilement être interrompu quant on n'est pas chez soi.

Règle C



On identifie ici aisément les axes F1 et F2. L'axe F2 oppose les lieux de lecture fixe au lieu de lecture plus instable comme expliqué précédemment. L'axe F1 illustre d'un coté la lecture dans un cadre personnel et d'un autre coté dans un cadre extérieur.

La contribution des variables « parents, Amis » et « Transports » à l'axe F2 sont respectivement d'un peu plus de 50% et de près de 32% ce sont donc évidement les variables les plus représentative de cette axe. Le fait que l'axe F2 illustre une pratique de lecture dans un lieu « instable » est donc indiscutable.

Sur l'axe f1 on observe que l'info régionale se situe plutôt à gauche et l'info nationale plutôt à droite.

L'info régionale est surtout lue chez les parents et amis et à domicile alors que les quotidiens nationaux sont lus soit à domicile soit sur le lieu de travail. Quant aux quotidiens urbains gratuits ils sont lus majoritairement dans les transports ou ils sont distribués (bouche de métro, arrêt de bus) et sur le lieu de travail ou on arrive après son trajet en transport en commun le matin. Encore une fois si Direct Soir était intégré à cette enquête on n'observerait plus les mêmes tendances, on aurait surement une lecture dans les transports et à domicile pour ce quotidien urbain gratuit.

D) Conclusion

Cette analyse fait ressortir plusieurs tendances. Les quotidiens urbains gratuits sont lus le matin dans les transports et sur les lieux de travail ce qui correspond à leur lieux et horaires de distribution.

L'information régionale est beaucoup lue à domicile et chez les parents et amis. Les quotidiens régionaux sont lus en milieu de matinée et sur l'heure de midi. Les hebdos régionaux sont plus lus l'après midi et le soir mais on peut se poser la question de la pertinence de l'analyse sur une journée de la lecture d'un hebdomadaire.

La presse nationale est lue en fin de matinée, sur l'heure de midi ainsi que le soir à domicile et au travail. Elle permet sans doute d'aborder l'actualité plus en profondeur une fois les brèves consultées dans les quotidiens gratuits et dans les quotidiens régionaux plus tôt dans la journée.

Cette analyse permet donc au final d'établir le profil type d'un lecteur de la presse quotidienne en France. Ces comportements peuvent aider à comprendre la segmentation de l'information entre les différents types de presse.

Fonctionnalité Web

A) *Présentation du tableau « Fonctionnalité web »*

1) *le tableau*

Par secteurs d'activités, en % des entreprises ayant un site	Présentation	Catalogue	Commandes en ligne	Paiement en ligne	Info client et suivi commande	SAV	Recueil de l'informations sur les clients	Diffusion et/ou recueil des offres d emplois
Services informatiques	89,3	63,6	14,9	0	4,6	20,7	26,6	66,9
Activités immobilières	87,2	65,9	9,9	1,8	8	9,5	13,8	20,5
Location sans opérateur	95,8	91,5	21,7	10,9	19,6	15,3	42,1	39,5
Télécommunications	100	88,7	26,1	16,5	28,4	33,2	38,2	71,3
Conseils et assistance hors services informatiques	97,1	60,2	11,6	7,8	3,9	6,8	19,1	32,6
Services opérationnels hors location	89,9	41,3	10,9	1,9	5,9	6,2	29,9	27,3
Hôtellerie et autres hébergements	99,1	93,6	55,7	25,2	23,1	10,8	28	18,3
Activités audiovisuelles	87,2	68,9	18,8	15,1	3,2	5,4	32	25,9

2) *Source*

Ce tableau est un résultat d'une enquête sur les Technologies de l'Information et de la Communication et le commerce électronique en 2002 du SESSI (Service des Etudes Statistiques Industriels), du SCEES (Service Central d'Etudes et d'Enquêtes Statistiques) et de l'INSEE (Institut National de la Statistiques et des Etudes Economiques). Ces différents organismes sont reconnus et il n'y a donc pas de doute quant à la fiabilité des données.

3) *Les individus*

Les individus de ce tableau sont des activités de services (comme les services informatiques, les activités immobilières, …). Pour le tableau nous avons pris en compte, dans chaque secteur, les entreprises de plus de 20 salariés qui possèdent un site internet.

4) *Les variables*

Les variables de ce tableau sont des fonctionnalités de site internet. Nous trouvons alors :

- Présentation : le site sert de présentation de l'entreprise et des ses activités
- Catalogue : le catalogue des produits et services est consultable via le site
- Commandes en ligne : il est possible de passer une commande directement par le site
- Paiement en ligne : le paiement peut être pris en charge directement sur le site (paiement sécurisé)
- Info client et suivi commande : chaque client dispose sur le site d'un profil et peut consulter à tout moment le suivi de ses commandes
- SAV : c'est le service après-vente, une aide en ligne est disponible aux clients
- Recueil de l'information sur les clients : le site recueil des données sur ses clients qui rentrent dans une base de données servant à la publicité adressé directement aux clients (par exemple des offres personnalisées)
- Diffusion et/ou recueil d'offre d'emploi : le site permet aux clients de consulter des offres d'emploi mais aussi de les diffuser.

5) *La problématique*

Il est donc intéressant d'étudier ce tableau pour déterminer l'utilisation de certaines fonctionnalités par rapport au type d'activité et si cela parait cohérent avec la réalité.

B) *Choix de la méthode*

Chaque méthode étudiée en analyse de données correspond à un type de tableau. En effet le tableau de mesure est associé à l'ACP et le tableau de contingence à l'AFC, par conséquent nous allons donc recourir ici à l'ACP.

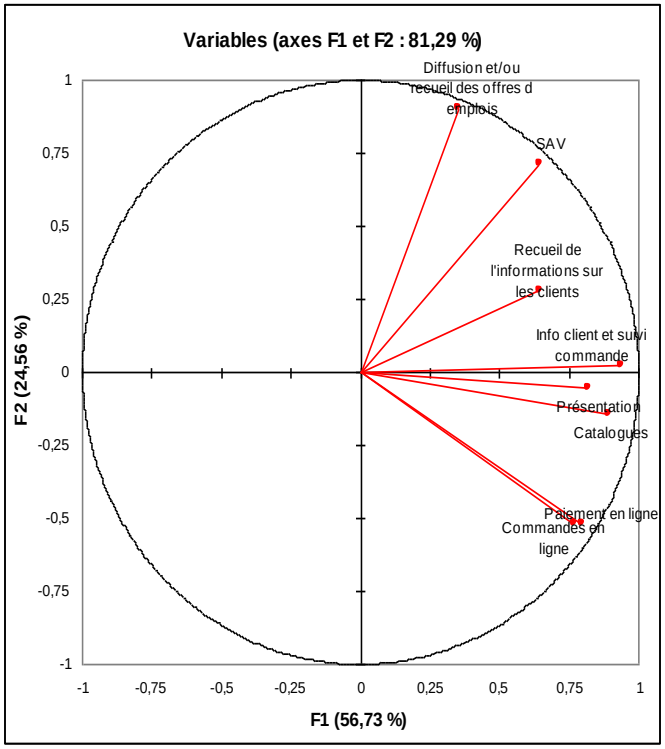
Les principes et objectifs de cette méthode ont été présentées dans la partie 1 de ce dossier.

C) *Analyse de l'ACP via XLstat*

Règle A

81,29% de l'information est réuni sur les axes F1 (56,73%) et F2 (24,56%). On peut observer dans les annexes la contribution de chaque axe à l'inertie, et l'on remarque qu'il y a un décroché entre l'axe F2 et l'axe F3. C'est pourquoi il n'est pas intéressant de retenir d'autres axes que les axes F1 et F2

Règle B



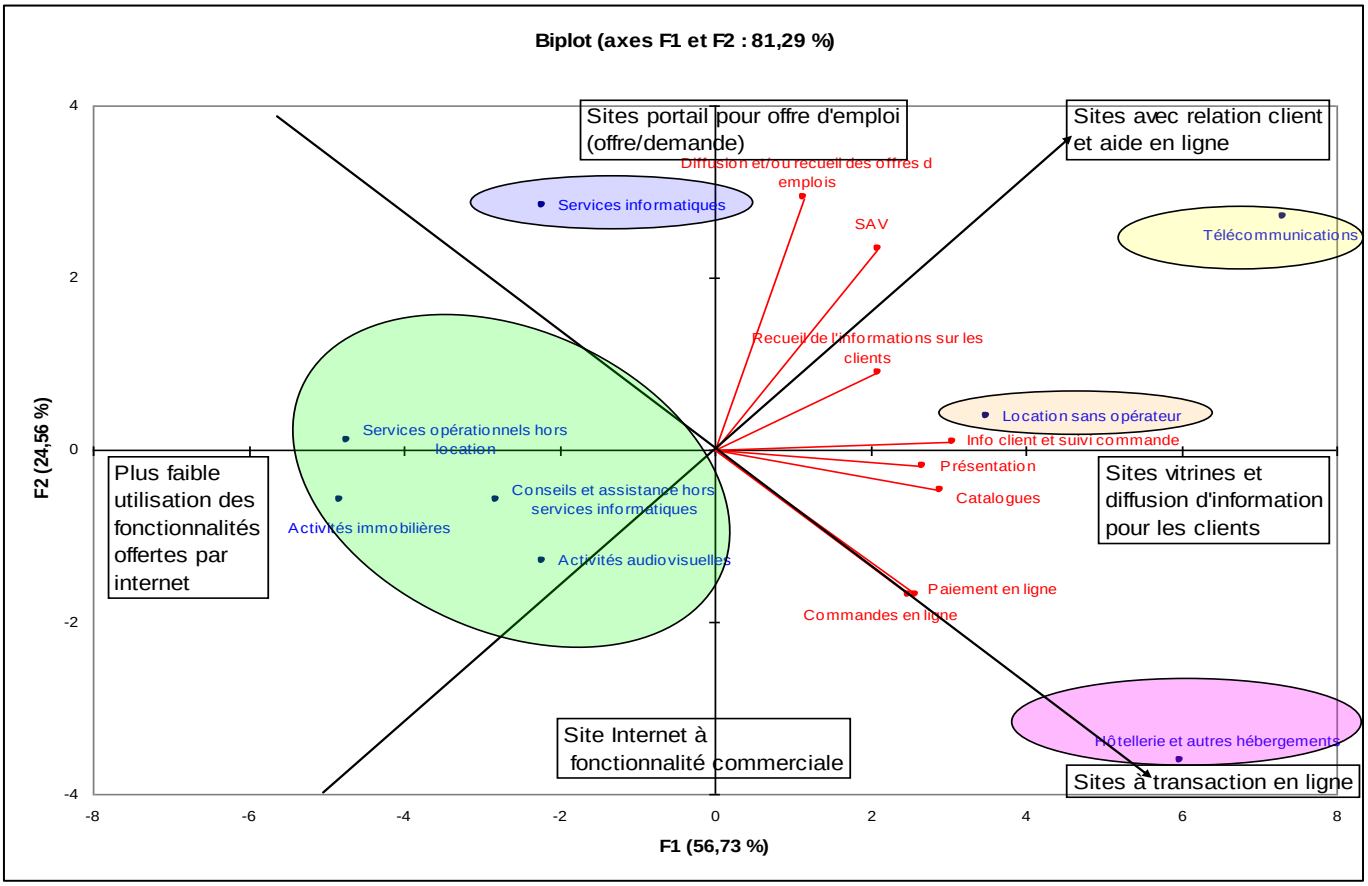
Le cercle des corrélations sert à définir les vecteurs des composantes des individus sur les axes principaux d'inertie. Les points du nuage des variables se trouvent importés à l'intérieur du cercle des corrélations. La corrélation est d'autant plus forte que les variables vont dans le sens de l'axe. Chaque variable est corrélée selon l'importance du taux d'utilisation des différentes fonctionnalités.

Plus les variables sont situées au bord du cercle, mieux elles sont représentées. Les corrélations des variables avec les axes favorisent la formation de composantes principales. Une composante principale est un *« paquet »* de variables très corrélées.

Dans le cadre de cette analyse, aucun vecteur ne joint le cercle. Par conséquent nous ne pourrons pas obtenir de conclusion parfaite. Plus les coordonnées des variables coïncident avec le cercle (se reprochent de 1), plus les interprétations sont pertinentes.

On peut déjà séparer les vecteurs en 2 grandes catégories sur l'axe F2. Les vecteurs *diffusion et/ou offre d'emplois*, *SAV*, *recueil de l'information sur les clients* et *information clients et suivi commandes* forment une catégorie de sites axés sur la relation client, tandis que les vecteurs *présentation*, *catalogues*, *paiement en ligne* et *commandes en lignes* forment une catégorie de sites plutôt commerciaux.

Règle C



1. Interprétations des axes

Pour bien déterminer l'interprétation des axes, il est nécessaire de voir la contribution des variables sur les axes F1 et F2.
La correspondance avec les axes peut paraitre évidente, et c'est pourquoi dans le cadre de cette étude j'ai choisi de compléter les axes F1 et F2 avec 2 axes : un axe Sud-Est et un axe Nord-Est.

- Axe Nord = Sites portail pour offre d'emploi (offre/demande)

La contribution de la variable *Diffusion et/ou recueil des offres d'emploi* sur l'axe F2 est de 41,395 % : c'est la variable la plus représentative de cet axe. Plus généralement cet axe nous aide à repérer les activités plus fortement utilisatrices de l'optique de la relation client au sein de leur site internet.

- Axe Sud = Sites Internet à fonctionnalité commerciale

Le regroupement des variables *'Commandes en ligne'*, *'Paiement en ligne'*, *'Catalogue'* et plus faiblement '*Présentation'*, nous permet d'avancer le fait que cette axe permet de déterminer les activités utilisant leur site surtout à but commercial.

- Axe Ouest = Plus faible utilisation des fonctionnalités offertes par Internet

Nous pouvons assez facilement dire que l'axe F1 est un axe croissant de l'utilisation des fonctionnalités, ce qui permet de déterminer les activités qui sont moins utilisatrices des différentes fonctionnalités offertes par Internet pour leur site web.

- Axe Est = Sites vitrines et diffusion d'information pour clients.

La contribution des variables '*Présentation*', '*Catalogue*' et '*Info client et suivi commande'*, respectivement de 14,877 %, 17,585 % et 19,499 % à l'axe F1 permet de déterminer cet axe comme celui indiquant les activités dont le site est surtout une vitrine pour les sociétés et permet de suivre l'état des commandes.

- Axe Nord-Est = Sites avec relation client et aide en ligne

Cet axe est la bissectrice des axes '*SAV*' et '*recueil de l'information sur les clients'*. Il permet donc de déterminer les activités qui utilisent leur site comme un outil d'aide en ligne avec un recueil des informations sur les clients, ce qui parait pertinent dans le cadre d'un service après-vente en ligne efficace.

- Axe Sud-Est = Sites à transaction en ligne

La forte corrélation entre les variables '*Paiement en ligne'* et '*Commandes en ligne'* nous permet de dégager une composante principale permettant de déterminer les activités qui utilisent plus les outils de transaction en ligne. Ce regroupement parait très logique car pour pouvoir payer en ligne, il faut passer une commande. La réalité nous le montre bien, il suffit pour cela de consulter les multiples sites commerciaux où l'on peut payer en ligne.

Nous voyons par ailleurs que cet axe et l'axe Nord-Est ne sont pas corrélés.

2. Groupement des individus

Groupe 1 = *Services opérationnels hors location, activités immobilières, conseils et assistance hors service informatique, activités audiovisuelles*

Ces activités de services sont moins utilisatrices des fonctionnalités commerciales et de relation clients que les autres que la moyenne. Ce fait peut paraitre logique lorsque l'on regarde de quel type d'activité il s'agit pour ces individus.

Groupe 2 = *Services informatiques*

Les sites des services informatiques proposent donc en particulier la diffusion et le recueil d'offres d'emploi, et dans une moindre mesure une aide en ligne avec un recueil d'information clients. Par contre les sites des services informatiques ne possèdent pas ou peu d'outils pour les transactions en ligne.

Groupe 3 = *Location sans opérateur*

Les sites des services de location sans opérateur tendent à proposer des outils de transaction en ligne ainsi que s'intéresser à la relation client, avec une proportion un peu plus grande pour cette dernière, et ce de façon supérieur à la moyenne. Par ailleurs les activités de location sans opérateur utilisent bien leur site comme une présentation, une vitrine des locations proposées.

Groupe 4 = *Télécommunications*

L'on remarque assez nettement dans les activités de télécommunications l'utilisation de leurs sites dans la relation client, l'aide en ligne et comme présentation des leur activité. Elles sont par contre seulement un peu plus utilisatrices des fonctionnalités de transaction en ligne que la moyenne.

Groupe 5 = *Hôtellerie et autres hébergements*

Les sites de ces activités sont beaucoup plus utilisatrices des transactions en ligne que les autres activités. Ces sites d'Hôtellerie et autres hébergements servent aussi beaucoup à présenter l'activité, les produits proposés ainsi que des informations sur le suivi des commandes.

D) Vérifications des interprétations

Groupe 1 : Si l'on prend par exemple les 'services opérationnelles hors location', leurs sites sont ceux qui proposent le moins des catalogues (41,3%), l'utilisation de la présentation (89,9%) est inférieur à la moyenne (93,2%), tout comme les autres variables. Il en est pratiquement de même pour les autres individus de ce groupe.

Groupe 2 : On voit déjà que les services informatiques n'utilisent pas le paiement en ligne, et les variables commerciales sont toutes inférieurs à la moyenne, tandis qu'à part pour « l'information client et le suivi de commande », les variables d'aide en ligne, relation clients et diffusion d'offre d'emploi sont supérieur à la moyenne, et l'on note par ailleurs le taux élevé dans cette dernière catégorie de 66,9% alors que la moyenne est à 37,78%.

Groupe 3 : Les taux des activités de location sans opérateurs sont au-dessus mais proche de la moyenne dans les fonctionnalités commerciales, mais supérieur à la moyenne dans l'aspect relation client, et même le taux de site pour le recueil de l'information sur les clients est le plus élevés des individus.

Groupe 4 : On remarque que dans les sites des activités de télécommunications, les taux d'utilisations de chaque fonctionnalité sont supérieurs à la moyenne, et que ce taux est le plus élevé pour la présentation, l'info client et suivi de commande, SAV et diffusion et/ou recueil d'offres d'emploi. On voit donc la spécialisation de cette activité dans des sites misant plus sur la relation client.

Groupe 5 : On remarque déjà que les taux d'utilisation des fonctionnalités commerciales sont tous supérieurs à la moyenne, et même les plus élevés pour 'catalogue', 'commande en ligne' et 'paiement en ligne'. Quant aux fonctionnalités de la relation client, seul le taux d'utilisation de la fonctionnalité 'Info client et suivi de commande' est supérieur à la moyenne, les autres étant inférieurs à la moyenne, sauf le 'recueil d'information sur les clients', légèrement supérieur à la moyenne.

E) Conclusion

On a pu déterminer grâce à cette analyse que les sites web de différentes activités de services utilisent des fonctionnalités différentes, que l'on peut regrouper principalement en 2 catégories : les fonctionnalités commerciales et les fonctionnalités de relation clients.

On remarque par ailleurs que le taux d'utilisation de telle ou telle fonctionnalité dans les sites internet dépend du type d'activité exercée par les entreprises de services.

Par ailleurs nous distinguons essentiellement 3 types d'activités plus fortes utilisatrices des fonctionnalités commerciales ou de relation client : les entreprises de télécommunications, d'hôtellerie et autres hébergements, et de location sans opérateur.

Nombre de Visiteurs par site internet

A) *Présentation du tableau « nombre de visiteurs par site internet »*

1) *le tableau*

	Europe 1	Jean Marc Morandini.com	Virgin Radio	Virgin 17	Mcm	Choc	Entrevue	Première	Virgin Mega.fr
15-24 ans	136320	118200	340990	71249	343672	94514	457693	87220	256578
25-34 ans	247080	190800	232349	43833	139536	70650	482754	211197	211878
35-49 ans	288615	203400	148291	22793	123386	105504	306008	230510	311112
>50 ans	391920	87600	70577	21518	38760	43018	71226	93450	114432
Total	**1063935**	**600000**	**792207**	**159393**	**645354**	**313686**	**1317681**	**622377**	**894000**

	Elle	Genealogie.com	Routard.com	Auto Moto.com	RFM	Paris Match	Tele7.fr	Vodeo TV	**Total**
15-24 ans	278694	71520	117135	10595	25058	18190	218436	39144	**1906436**
25-34 ans	270357	77480	306090	11895	52921	32470	280560	32760	**1830077**
35-49 ans	400176	184760	196650	16705	84337	60010	288576	51072	**1739619**
>50 ans	241773	262240	235125	25805	24684	59160	214428	44856	**932501**
Total	**1191000**	**596000**	**855000**	**65000**	**187000**	**169830**	**1002000**	**167832**	**6408633**

2) *Source*

Ce tableau est le résultat de différentes études. Nous avons pris soin de le transformer, les données de profils étant en pourcentage et les totaux représentant le nombre de visiteurs par site. Provenant du site www.lagardere-active-pub.com, site web du géant français de la communication et des médias, les résultats sont très peu contestables. En effet, les sources concernant les profils proviennent de l'institut de sondage IPSOS Profiling, et ont été publiés récemment (Oct-Nov 07). La réputation du groupe n'est plus à faire. D'autre part, le nombre de visiteurs unique par site a été publié en février 2008 par le cabinet Nielsen NetRatings.

Ce cabinet utilise une technologie capable de mesurer les usages d'Internet pour fournir au marché mondial et aux marchés locaux l'information la plus pertinente et compréhensible sur de multiples indicateurs (nombre de visiteurs uniques, profil des visiteurs, taux de pénétration, nombre de visites, nombre de pages vues, durée de connexion, …). Disposant d'une technologie de pointe, Nielsen NetRatings est à même de livrer dans ce domaine une mesure fiable (50 années d'expérience dans la mesure d'audience).

3) Les individus

Les individus représentés dans ce tableau correspondent à des sites web. Nous pouvons y trouver les sites suivants :

- **Europe 1 :** Site où l'on traite de l'actualité française et internationale et où l'on peut écouter la web radio d'Europe 1.
- **Jean Marc Morandini.com :** site web consacré à l'actualité des medias.
- **Virgin Radio :** Site de web radio.
- **Virgin 17 :** site web de la chaîne de télévision musicale française privée à caractère commercial.
- **Mcm :** Site web de la chaine musicale.
- **Choc :** Site web de zapping du net, photos et vidéos insolites.
- **Entrevue :** Site web ou l'on traite de l'actualité people, de l'actualité des stars et de vidéos insolites.
- **Première :** On peut retrouver sur ce site toute l'actualité du cinéma.
- **Virgin Mega.fr :** Boutique de vente de disques et de titres en téléchargement de Virgin en ligne.
- **Elle :** Magazine de mode pour les femmes en ligne.
- **Genealogie.com :** Il s'agit d'un portail de la généalogie en France pour consulter son état-civil en ligne, construire son arbre généalogique,…
- **Routard.com :** On peut trouver sur ce site des informations pour préparer son voyage, un guide des hôtels et restaurants en France.
- **Auto Moto.com :** Magazine sur l'actualité automobile en ligne.
- **RFM :** Site de la web radio RFM .
- **Paris Match :** Magazine d'actualité et people en ligne.
- **Tele7.fr :** Guide des programmes TV.
- **Vodeo TV :** Site internet de vidéo à la demande (VOD) édité par La Banque audiovisuelle.

4) Les variables

Les variables du tableau correspondent aux tranches d'âges des visiteurs des sites internet. Les âges mis en classe ici sont les suivants : 15-24 ans, 25-34 ans, 35-49 ans et les plus de 50 ans. L'âge par nature est quantitatif, cependant, étant donné que la variable a été classée, cela équivaut à un caractère qualitatif.

5) La problématique :

L'intérêt de cette analyse est ici d'identifier la segmentation des visiteurs de différents sites web. Il est intéressant de voir ici les liens entre les différentes modalités présentes dans ce tableau, afin d'en extraire les correspondances, les tendances et les profils moyen de celles-ci.

B) Choix de la Méthode

Comme nous avons pu le voir précédemment, l'analyse factorielle des correspondances est utilisée pour synthétiser une information importante en volume, contenue dans un tableau de contingence. C'est le cas de notre tableau, il est donc utile d'utiliser cette méthode. En effet, nous retrouvons dans notre tableau n visiteurs, et le contenu de la case située en ligne i et en colonne j correspond à l'effectif n_{ij} (nombre de visiteur i en j).

C) Analyse de l'AFC via XLstat

Règle A

Afin d'établir la part de l'information expliquée par le plan, la première règle est essentielle. Elle va nous permettre d'analyser notre tableau sans problème. Nous nous intéresserons donc ici au test du Khi², ainsi qu'au tableau des valeurs propres, et les pourcentages d'inertie.

Le plan représenté par le graphique symétrique restitue 91,04% d'inertie, ce qui signifie que l'information est ici y très bien restituée. En effet, un minimum de 60 % de l'information doit être contenu par l'ensemble des axes pour pouvoir y faire une analyse solide. Le premier axe restitue 74,80% de l'information, quand au second, la part présente est de l'ordre de 16,23%. Il est donc inutile d'intégrer dans cette analyse un troisième axe, la part relative à l'information étant très faible (8,9%). Les deux axes principaux restitués seront ainsi les axes F1 et F2.

Le test du Khi² s'intéresse à la différence entre la valeur observée (Oij) et la valeur attendue (Eij) (ou valeur théorique) s'il y avait indépendance.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad E_{i,j} = \frac{O_{i+} \times O_{+j}}{N}$$

Afin d'interpréter le test du Khi², nous nous retrouvons face à deux hypothèses :

H0 : Les lignes et les colonnes du tableau sont indépendantes.

Ha : Il existe un lien entre les lignes et les colonnes du tableau.

Etant donné que la p-value calculée est inférieure au niveau de signification alpha=0,05 (<0,0001), on doit rejeter l'hypothèse nulle H0, et retenir l'hypothèse alternative Ha. De plus, Il est à noter que le risque de rejeter l'hypothèse nulle H0 alors qu'elle est vraie est inférieur à 0,01%. Nous pouvons donc conclure que les lignes et les colonnes de notre tableau de contingence sont dépendantes entre elles.

Règle B

La seconde règle peut être divisée en deux points :

- Premièrement, on va mesurer la dépendance entre les caractères en fonction de l'éloignement des modalités du centre.
- Ensuite, nous allons repérer les modalités ayant le profil moyen, à savoir celles qui sont situées près du centre.

Dépendance entre les caractères :

L'éloignement d'une modalité (ligne i ou colonne j) au centre (origine des axes) reflète la différence du profil correspondant (distribution conditionnelle) au profil moyen (distribution marginale). Les modalités les plus éloignées du centre vont être originales et vont donc créer de la dépendance entre les caractères.

Dans notre graphique, on peut voir que plusieurs de nos modalités sont éloignées du centre. Les modalités Mcm, Virgin 17, et Virgin radio sont en effet très éloignés du centre. Elles sont également très proche de la modalité 15-24 ans. Il apparait ainsi une dépendance entre ces trois sites web, et les visiteurs âgés de 15 à 24 ans.

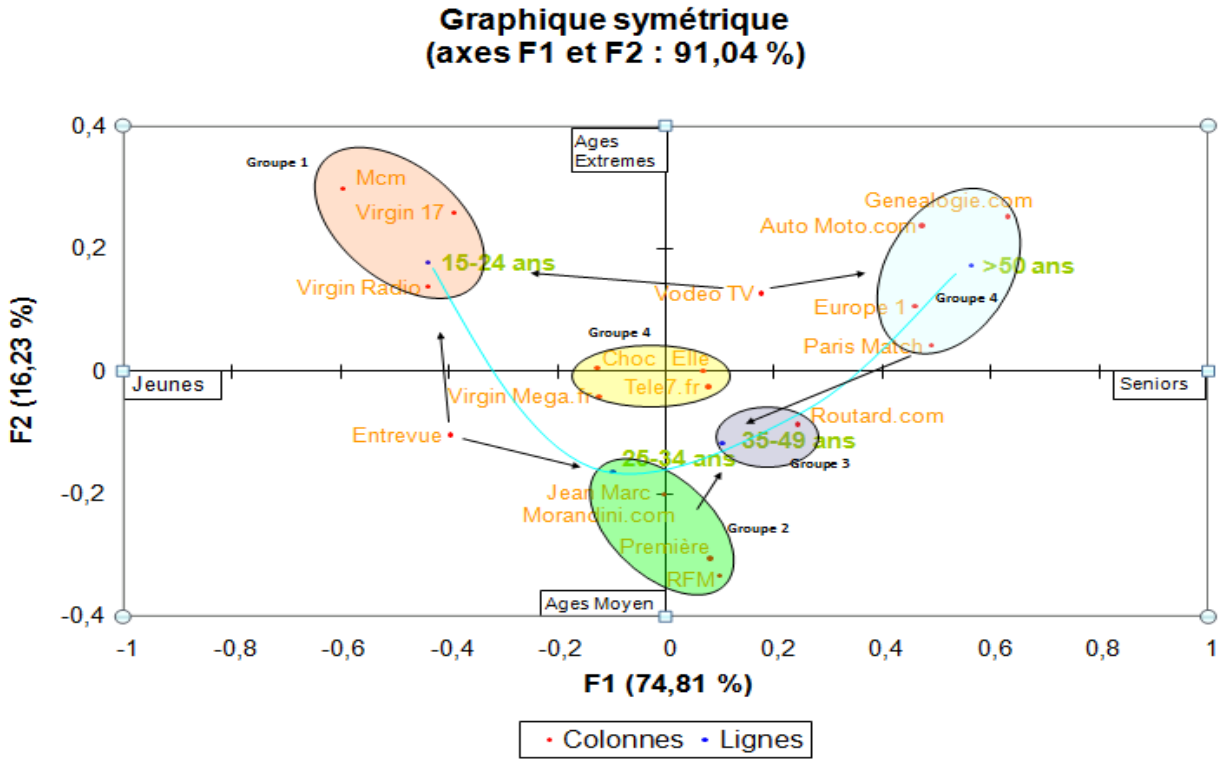
Dans la partie Est du graphique, on peut apercevoir également un éloignement de modalités. Il s'agit des modalités genealogie.com, auto moto.com, Europe 1. Ces sites web étant proche de la tranche d'âge des plus de 50 ans, la dépendance peut être établie.

Les dernières modalités très éloignés du centre sont les suivantes : Jean Marc Morandini.com, Premiere, RFM. Ces modalités apparaissent comme étant liés à celle des 25-34 ans car elle sont très proche géographiquement.

Profil moyen des modalités :

Les modalités ayant le profil moyen sont situées près de l'origine. Ainsi, on peut voir rapidement celles dont la dépendance n'est pas présente. Celles représentées près du centre sont ici : Elle, Choc, Télé 7.fr et Virgin Mega.fr. Elles ont donc sensiblement le profil moyen. Elles ne sont pas dépendantes d'une tranche d'âge particulière. Ainsi l'ensemble des visiteurs (toutes tranches d'âges comprises) vont sur ces sites de la même façon. Le tableau des profils lignes nous révèle en effet le faible écart entre les fréquences de ces modalités et la moyenne.

Règle C



Interprétation des Axes :

Comme nous avons pu le voir précédemment, il est important de regarder la contribution des variables sur les axes F1 et F2. On pourra donc établir certaines correspondances. L'interprétation des axes ici, va se faire selon 4 axes : Axe Nord, axe Sud, axe Ouest, axe Est.

- Axe Nord = Ages Extrêmes

Les contributions à l'axe F2 importantes concernant les individus sont ici Mcm et Genealogie.com. Ces dernières contribuent à l'axe F2, respectivement de l'ordre de 21,50%, et 14,30%. Elles correspondent aux âges extrêmes. En effet, Ils y a de fortes correspondances entre la modalité Mcm et la catégorie des 15-24 ans, et entre la modalité Genealogie.com avec les plus de 50 ans.

- Axe Sud = Ages Moyen

Les modalités Jean Marc Morandini.com, Premiere et RFM, sont très représentées dans cet axe. En effet, Premiere contribue de façon significative avec une part allant jusqu'à 22%. Cela nous permet ainsi de dégager l'axe F2, avec le regroupement des modalités très proche du sud, avec les modalités de classes d'âge moyennes.

- Axe Ouest = Jeunes

Les variables extrêmes sont très représentées dans l'axe F1. Ici l'axe Ouest est représenté très clairement par les Jeunes. Les 15-24 ans contribuent de l'ordre de 42,20% pour l'axe F1 et sont donc très significatifs. La modalité présente sur cet axe est donc très corrélée aux autres modalités du même groupe.

- Axe Est = Seniors

Quand à l'axe Est, il est représenté par les Seniors. Les plus de 50 ans participent à l'axe de façon très importante. Avec 52,9% de contribution à l'axe F1, la dépendance avec les modalités du même groupe y est très élevé. L'axe F1est donc très significativement une fonction croissante de l'âge des visiteurs des sites web enregistrés. Ces modalités 15-24 ans et > 50 ans se détachent très largement du profil moyen.

Groupement des individus :

Nous avons pu ici regrouper les modalités en 5 groupes distincts, et certaines tendances.

Groupe 1 : Mcm, Virgin 17, Virgin Radio et les 15-24 ans

On peut remarquer que les sites web présents dans ce groupe sont très éloignés de l'origine et par conséquent très dépendant de la classe d'âge 15-24 ans, celle-ci étant proche géographiquement. Ils sont assimilés à une population relativement jeune, en cohérence avec la nature de ces sites.

Groupe 2 : Jean Marc Morandini.com, Premiere, RFM et les 25-34 ans

Le groupe 2 est très proche de l'axe Sud, ce qui signifie qu'ils sont fortement corrélés aux âges moyens. Il existe cependant une dépendance plus forte de ces sites avec la catégorie des 25-34 ans. Quand au site Jean Marc Morandini.com, il est visité de la même façon pour les 25-34 ans et les 35-49 ans.

Groupe 3 : Routard.com et les 35-49 ans

Le site Routard.com possède une particularité. En effet, il correspond un peu au profil moyen, celui-ci n'étant pas très éloigné du centre. Il est quasiment visité de la même façon par les 25-24 ans et les plus de 50 ans. C'est pour cela qu'il se trouve dans le groupe des 35-49 ans, qui se retrouve lui aussi, guère loin du profil moyen.

Groupe 4 : Genealogie.com, Auto moto.com, Europe 1, Paris Match et les plus de 50 ans

Ce regroupement de sites web est très dépendant de la tranche d'âge des plus de 50 ans. Très éloignés du centre et proche de cette catégorie, ils sont donc assimilés au groupe des seniors. Paris Match se démarque un peu par une tendance vers le groupe des 35-49 ans.

Groupe 5 : Choc, Elle, Télé 7.fr et Virgin Mega.fr

Ce groupe correspond au profil moyen. Ces sites web sont visités de la même façon par toutes les catégories d'âges, et se retrouvent par conséquent très proche de l'origine.

Les autres tendances : Entrevue, Vodeo TV

Entrevue ne se retrouve dans aucun groupe distinct. Cependant, il est visité de la même façon pour les 15-24 ans et les 25-34 ans. Il correspond donc à une classe de visiteurs relativement jeune. Quand à Vodeo TV, celui-ci est aussi bien visité par les 15-24 ans que par les plus de 50 ans, se retrouvant ainsi, proche de l'axe Nord concernant la catégorie d'âges extrêmes.

D) Vérifications et interprétation

Groupe 1 : En regardant les tableaux des profils lignes et colonnes, on arrive à une conclusion similaire. La part des 15-24 ans concernant les sites Virgin Radio, Virgin 17, et Mcm, sont respectivement de 43,04%, 44,70% et de 53,25%. En observant le tableau des profils colonnes, on observe une distribution marginale supérieur à la moyenne pour l'ensemble de ces trois sites (12,70%, 2,65% et 12,80%) et supérieur au profil moyen observé.

Groupe 2 : Le groupe des 25-34 ans et des 35-49 ans étant proche, on peut voir des similarités dans les dépendances. En effet, la part de ces deux tranches d'âges est quasiment identique pour les trois sites web, en regardant le tableau des profils lignes. Le tableau des profils colonnes nous montre également que ces modalités sont au dessus du profil moyen, et sont très liés à la catégorie des âges moyens. Ce groupe est donc en relation très forte avec la tranche des 35-49 ans.

Groupe 3 : En regardant le tableau des profils colonnes, on remarque que le groupe des 35-49 ans est très proche du profil moyen, d'où sa positon. Concernant le site Routard.com, il est au dessus du profil moyen pour les tranches d'âges des plus de 50 ans et des 25-34 ans, et de même très proche en termes de fréquence. Le site est donc visité semblablement par ces deux tranches d'âge.

Groupe 4 : On remarque que la part des plus de 50 ans observée dans le tableau des profils lignes est très importante pour l'ensemble des sites du groupe. De plus, dans le tableau des profils colonnes, on peut voir que les fréquences sont au dessus du profil moyen pour ces quatre sites. Cependant la correspondance entre les plus de 50 ans et la modalité Paris Match est un peu moins significative.

Groupe 5 : Ce groupe correspond au profil moyen. Le tableau des profils lignes conforte cette observation, les distributions sont en effet très proche de celle du profil moyen. Il en est de même pour le tableau des profils colonnes.

Entrevue, Vodeo TV : Pour le site web Entrevue, on observe dans les tableaux de fréquences la même similarité : Les fréquences sont au dessus des profils moyen concernant les tranches d'âges de 15 à 24 ans et de 25 à 34 ans (respectivement 17,04% et 16,68% pour un profil moyen de 12,38% dans le tableau des profils colonnes). On remarque que le site de Vodeo TV est visité de la même façon pour les 15-24 ans et les plus de 50 ans, cependant il est plus visités par les 25-34 ans et les 35-49 ans, d'où sa position dans le graphique.

E) Conclusion

Grâce à l'analyse factorielle des correspondances, nous avons pu établir des liens entre la catégorie d'âge des visiteurs de sites web et les sites web en question. Cela permet ainsi aux sites d'effectuer une segmentation selon l'âge de leurs visiteurs. Il est à souligner qu'il apparait dans notre graphique un « effet Guttman ». En effet, nos modalités sont ordonnées selon l'âge. En joignant les points qui les représentent, on peut voir apparaitre une parabole indiquant que sur un axe, l'ordre est respecté, allant des jeunes eux seniors, mais sur l'autre axe, il y a opposition entre les modalités d'âge extrêmes et d'âge moyens.

CONCLUSION

Les outils que l'analyse de données met à notre disposition, tels que l'Analyse en Composantes Principales (ACP) et l'Analyse factorielle des Correspondance (AFC) nous ont permis d'effectuer nos analyses et de pouvoir en tirer des conclusions.

Nous avons pu ainsi déterminer des comportements dans les habitudes de lectures de la presse, au niveau des lieux et des horaires.

Nous avons également mis en évidence que les fonctionnalités des sites Internet d'entreprise dans des activités de services sont différentes suivant le type d'activité.

Enfin les sites Internet différents que nous trouvons sur la toile attire des publics différents, des plus jeunes aux plus âgés, ce qui rejoins souvent le public visé par ses sites.