UNIVERSITÉ DE MONTRÉAL

CLASSIFICATION DE MOTS-CLÉS DES CAMPAGNES PUBLICITAIRES SUR LES MOTEURS DE RECHERCHE ET CALCUL DE PRÉVISIONS

CHAKIR ASSARI

DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION

DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES

(GÉNIE INDUSTRIEL)

AOÛT 2014

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

CLASSIFICATION DE MOTS-CLÉS DES CAMPAGNES PUBLICITAIRES SUR LES MOTEURS DE RECHERCHE ET CALCUL DE PRÉVISIONS

présenté par : ASSARI Chakir

en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées

a été dûment accepté par le jury d'examen constitué de :

M. AGARD Bruno, Doct., président

M. GAMACHE Michel, Ph.D., membre et directeur de recherche

M. ADJENGUE Luc, Ph.D., membre et codirecteur de recherche

M. PARTOVI NIA Vahid, Doct., membre

DÉDICACE

Je dédie ce mémoire aux deux êtres les plus chers à mon cœur, ma mère et mon père qui sont l'essence de mon existence.

REMERCIEMENTS

Tout d'abord, je remercie grandement mon directeur de recherche, Michel Gamache, qui m'a proposé ce thème et qui m'a guidé le long de sa réalisation.

Je ne saurais être reconnaissant envers lui pour sa patience et de son soutien, c'est pourquoi je lui suis toujours redevable.

Je tiens aussi à remercier mon co-directeur, Luc Adjengue, pour m'avoir accordé son aide, son expertise, son conseil et sa disponibilité durant mon projet.

Je remercie également mes deux professeurs, Bruno Agard et Vahid Partovi Nia, qui m'ont enseigné les méthodes et les techniques de Data Mining. Sans oublier leurs rencontres organisées qui m'ont beaucoup aidé dans ma recherche.

J'exprime pareillement ma gratitude envers Sandrine Paroz de la compagnie d'Acquisio, qui m'a orienté et fourni les données nécessaires à la réalisation de ce projet. Sans bien sûr oublier Mohammed et Younes d'Acquisio qui m'ont initié le code nécessaire à l'apprentissage du langage SQL. Tous mes remerciements à l'équipe d'Acquisio pour m'avoir accordé leur temps nécessaire à la réalisation de mon projet.

Je remercie mon ami le linguiste Mounir Zaidi et Riad Kechroud, qui m'ont aidé dans la rédaction de ce mémoire.

Je conclue les expressions de ma profonde gratitude en remerciant ma famille, mon épouse et mes deux enfants, Imène et Ayoub, qui ont toujours été là pour me procurer la joie de vivre et être ma flamme qui me donne davantage l'énergie suffisante me laissant toujours aller de l'avant.

RÉSUMÉ

Les publicités sur le web qui se présentent sous forme de liens textuels s'affichent dans les pages de résultats des moteurs de recherche, suite aux requêtes des internautes par le biais des mots-clés achetés par les annonceurs via un système d'enchères. Communément, les premières pages du moteur de recherche offrent aux annonceurs de promouvoir les produits et services. Quand une annonce s'affiche et qu'un internaute clique sur le lien correspondant, l'entreprise en question paie le moteur de recherche. Afin de gérer son budget, une entreprise doit établir des stratégies d'enchères; sélectionner un ensemble de mots-clés et déterminer un montant pour chaque mot-clé. Les données historiques de ces mots-clés sont évidemment nécessaires pour évaluer le comportement des requêtes des internautes sur les moteurs de recherches.

Les travaux présentés dans ce mémoire sont une suite d'un thème de recherche dont l'objectif est de développer des algorithmes permettant d'améliorer le rendement des campagnes publicitaires sur les moteurs de recherche. Dans cette optique un algorithme permettant d'affecter des positions optimales aux mots-clés est développé de sorte que le nombre total de clics par campagne est maximisé. En outre, une méthode de génération de courbes génériques est proposée pour chaque mot-clé afin d'effectuer une prédiction à la fois du nombre de clics et du coût par clic en fonction de sa position. Ces paramètres sont essentiels au programme d'optimisation.

Nous présentons une approche de classification basée sur les techniques de data mining pour l'extraction des connaissances cachées dans une base de données de mots-clés. Le but est de déceler des comportements similaires au niveau des mots-clés et de les classifier par la suite. Si la classification des mots-clés est optimale, on estime pouvoir obtenir des courbes génériques de meilleure qualité.

Notre stratégie utilise beaucoup d'échantillons composés de différentes campagnes dans différents types de marché sur le Web. Cette stratégie nous permet de conclure, lors de la classification automatique, que le nombre de classes de mots-clés est approprié pour toutes les campagnes publicitaires.

Nous exploitons divers méthodes de classifications automatiques pour une meilleure organisation des mots-clés selon leurs caractéristiques. Parmi les algorithmes cités dans le présent document il y a : k-means, fuzzy c-means, Clara, Clues et Pam.

Les résultats obtenus lors de la classification non supervisée se sont avérés en deçà de nos attentes. Toutefois, notre mandat ne s'arrête pas là, on doit améliorer les courbes génériques existantes.

Une évaluation expérimentale basée sur nos données montre que notre approche améliore modestement la précision des paramètres. Cependant, nous n'affirmons pas nécessairement que les résultats ainsi obtenus soient concrets car aucune de nos expériences pratiques n'a été conduite en temps réel sous le moteur de recherche Google.

ABSTRACT

Web advertisements that are in the form of text links are displayed in the results pages of search engines through internauts requests via keywords purchased by advertisers via an auction system. Commonly, the first pages of search engine offer advertisers to promote products and services. When an ad displays and a user clicks on it, the company in question pays the search engine. To manage its budget, a company must establish bid strategies; select a set of keywords and determine an amount for each keyword. Obviously, historical data on such keywords are needed to assess the behavior of users by their entry into the search engine query.

The work presented in this thesis is part of a series of research aimed at developing algorithms to improve the performance of advertising campaigns on search engines. In this context, we propose an algorithm that assigns optimal keywords positions so that the total number of clicks per campaign is maximized. Furthermore, a generic method of generating curves is proposed for each keyword to make a prediction of the number of clicks and estimate the cost per click according to its position. These parameters are critical to the optimization program.

We present a classification approach, based on data mining techniques, in order to extract hidden information in data warehouse keywords in order to identify similar behaviors of keywords and classify them thereafter. An improved classification of keywords is estimated to lead to better generic curves.

Our strategy uses a lot of samples from different campaigns with different types of market on the web. This strategy allows us to conclude, during the automatic classification, that the number of classes of keywords is appropriate for all campaigns.

We used various methods of automatic classifications for better organization of keywords according to their characteristics. Among the algorithms mentioned in this document there are: k-means, fuzzy c-means, clustering large application, clustering based on local shrinking and partitioning around medoïds.

The results obtained in the automatic classification proved to lower our expectations. However, our mandate does not stop there; we must improve existing generic curves.

Experimental evaluation based on data provided showed that our approach modestly improves the accuracy of parameters. However, we cannot say that the results are real because we have not done a practical experience in real time on the Google search engine.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

AdWords	« Advertiser Words » : expression anglaise utilisée pour désigner une annonce textuelle
CPC	Coût par clic
CTR	« Click Through Rate » : expression anglaise utilisée pour désigner le taux de clics
ECD	Processus d'extraction de connaissances à partir des données
MAE	« Mean Absolute Error » expression anglaise pour désigner une mesure d'erreur.
PAM	« Partitioning Around Medoïds » expression anglaise pour designer le nom d'un algorithme de classification non automatique.
SERP	« Search Engine Result Pages » : expression anglaise pour désigner la liste de résultat de recherche après une requête
URL	« Uniform Resource Locator » : expression anglaise pour désigner le localisateur uniforme de ressources, appelé aussi adresse Web.

LISTE DES ANNEXES

CHAPITRE 1 : INTRODUCTION

Le e-commerce représente un chiffre d'affaire de plusieurs milliards de dollars, son taux de croissance ne cesse d'augmenter durant la dernière décennie. De nos jours, les agences publicitaires transitent systématiquement par les moteurs de recherche, tels que Google, Bing, Yahoo afin de promouvoir leurs annonces où chaque agence veut que son produit soit à la une. À cet effet, Google a établi un système d'enchères en temps réel de mots-clés que les internautes utilisent pour leurs recherches. Ce système positionne les liens des sites des clients sur la page du moteur de recherche en question. Quotidiennement, Google envoie les statistiques relatives aux mots-clés aux agences qui se basent sur celles-ci pour gérer leurs budgets et établir des stratégies d'enchères. Toutes ces agences publicitaires dépendent autour des fonctionnalités du texte. Il est difficile pour eux d'interpréter l'intention de rechercher des utilisateurs potentiels que par les mots-clés de la requête. Des plateformes de solutions de supports sont nées et proposées par des agences spécialisées en gestion de campagnes publicitaires sur les moteurs de recherche. Celles-ci font appel aux méthodes mathématiques et statistiques afin de gérer les campagnes de mots-clés. À titre d'exemple, Acquisio est le chef de la file mondiale des plateformes de solutions médias. Leur plateforme aide à effectuer le suivi, à gérer, à optimiser et à produire des rapports. Acquisio gère plus d'un demi-milliard de dollars publicitaires.

Plusieurs algorithmes d'optimisation et de prédiction sont appliqués par ces plateformes afin d'améliorer le rendement des campagnes publicitaires, en affectant les annonces aux diverses positions selon le budget attribué à ces dernières. Ces systèmes automates se basent sur les paramètres estimés dont les prédictions reposent sur l'hypothèse que les taux de décroissance relatifs du nombre de clics et du coût par clic sont relativement constants d'un mot-clé à l'autre (Quinn, 2011).

Ce mémoire a pour but de parvenir à une catégorisation des mots-clés afin d'améliorer les résultats de recherches obtenus par la méthode de Quinn (2011). On reviendra sur cela après avoir défini quelques termes propres à ce domaine qui permettront de mieux définir le problème.

1.1 Définition et concept

Afin de mettre à l'aise le lecteur le long de ce mémoire et pour éviter de le référer à toutes les fois aux mémoires faits auparavant, quand il s'agit d'un terme technique ou d'expressions propres au domaine des gestions de campagnes des annonces publicitaires sur les moteurs de recherche, nous proposons la définition de certains termes et concepts essentiels.

Moteur de recherche

Un moteur de recherche est une application informatique exécutée sur internet permettant de retrouver facilement des sites Web, des images, des vidéos à partir des mots saisis par l'utilisateur.

Il existe plusieurs moteurs de recherche sur internet sous des interfaces différentes, citons entre autres Google, Yahoo, Bing, Baidu, Yandex. Leurs outils de recherche sont des programmes sous forme de scripts appelés robots qui explorent l'ensemble des sites Web par leurs liens URL pour découvrir des nouvelles pages dans le but de les indexer et enregistrer dans sa propre base de données (serveurs d'index). Le parcours des pages web se fait périodiquement afin de constater les changements qui ont eu lieu. Chaque moteur utilise son propre robot dans le but de maintenir son index à jour (Google, 2013).

Tandis que les internautes saisissent les mots qui composent les expressions de leurs requêtes dans la zone texte d'un moteur de recherche, les résultats s'affichent, en général, de manière à ce qu'un certain nombre de liens commandités soient affichés en haut à gauche et / ou à droite de la page des résultats, communément appelées liens commerciaux. Les autres résultats sont positionnés en bas des liens commerciaux s'il y a lieu, sinon au-dessous de la zone texte du moteur. Ceux-là sont appelés résultats de recherche naturels ou Search Engine Result Pages (SERP) selon la version anglaise (Figure 1.1).

Lors de cette recherche, un algorithme est appliqué afin d'identifier les documents qui correspondent mieux aux mots contenus dans cette requête et de présenter les résultats recherchés par ordre de pertinence.

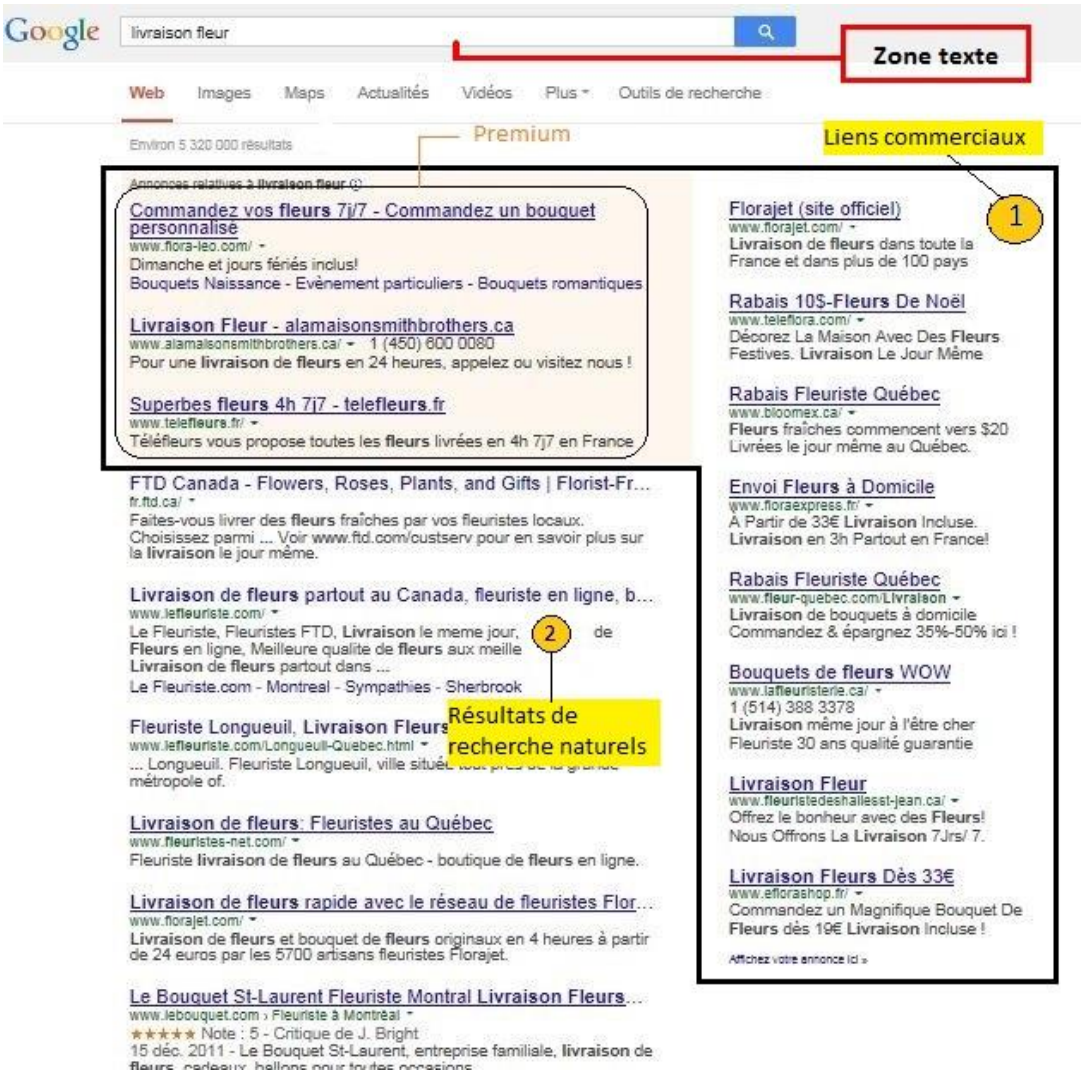


Figure 1.1: Exemple d'interface d'un moteur de recherche de Google

En résumé, le fonctionnement d'un moteur de recherche se fait en trois étapes principales:

- L'exploration
- L'indexation
- La requête de recherche

Il est important de présenter le principe derrière un moteur de recherche afin de préciser l'étape de notre étude dans le processus du moteur de recherche. Nous nous intéressons uniquement à la partie « requête de recherche du moteur de Google ». En effet, ces algorithmes font l'objet de très nombreux projets d'explorations scientifiques, et le moteur en question est le plus utilisé dans le monde.

Types de publicités

Le modèle économique le plus répandu du Web est la publicité en ligne. Les revenus générés par ce type de publicités sont considérables. À titre d'illustration, on estime qu'en 2013, Google en a récolté le tiers de l'ensemble, soit 38.6 milliards de dollars de revenu de publicité en ligne selon Statista (2013).

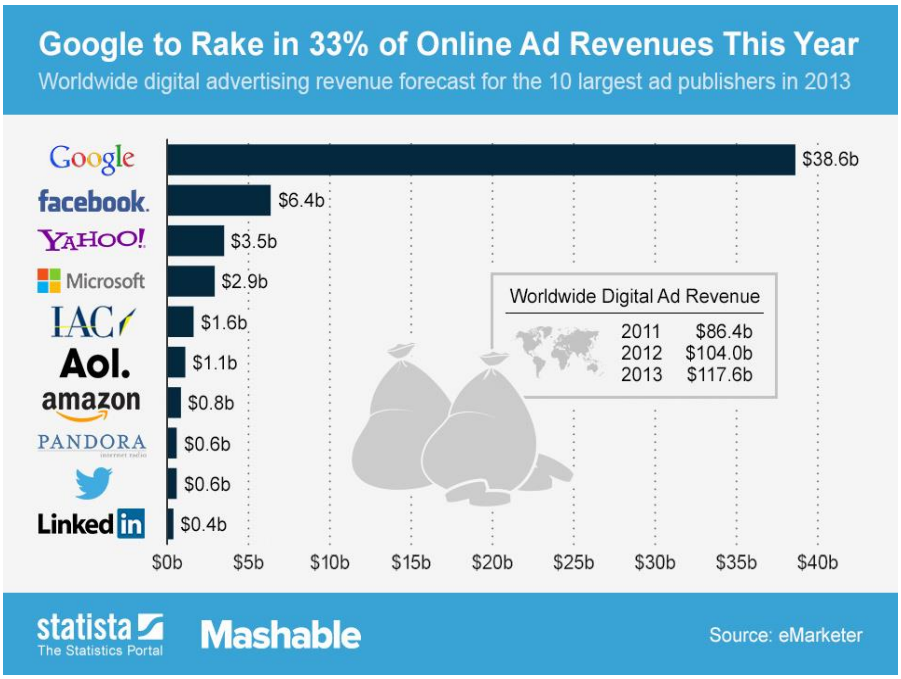


Figure 1.2: Revenus publicitaires en ligne (2013)

Il existe différents types de publicités sur internet tels que les bannières publicitaires, les annonces textuelles, les diffusions sur les médias sociaux, les publicités par courriel, etc.

Les deux principaux types de publicité sont les bannières et les annonces textuelles.

Bannière :

C'est une représentation graphique qui apparaît sur une page Web, comportant en général le nom et la nature du site Web de l'annonceur, associé au moyen d'un lien hypertexte. La bannière publicitaire peut également être accompagnée d'un contenu audio ou vidéo. La taille, la forme et l'emplacement de la bannière sur une page Web peuvent varier et sont déterminés en tenant compte du budget de l'annonceur.

Annonce textuelle :

Une annonce textuelle est constituée uniquement d'un texte qui décrit le produit ou le service offert. Elle est parfois plus appropriée que les autres types de publicité notamment lorsqu'elle est transmise à ceux qui utilisent des dispositifs mobiles. Le téléchargement d'une annonce textuelle se fait rapidement selon le Réseau Ontario des entreprises, affaires et économie (2014). Comparativement aux bannières publicitaires, les annonces textuelles sont efficaces pour réaliser des ventes. Elles ciblent mieux le besoin des internautes qui recherchent des informations en rapport avec les sociétés.

Le format des annonces textuelles est simple (Figure 1.3) et est composé de quatre lignes décrites comme suit:

- Un titre d'annonce
- Une adresse Web de l'annonceur (URL)
- Une description composée de deux lignes et dite « créatif » visant à donner plus de détails sur le produit ou service de l'annonceur.

Leur affichage est uniquement observé à la partie des résultats des liens commandités, il y a au maximum 11 liens par page dont trois liens au plus, dits « premium », illustrés ci-dessus (Figure 1.1). Il n'y a pas de distinction entre les positions Premium et standard en dehors de leur position dans le classement.

Les annonces textuelles sont appelées « AdWords » d'après Google.

Figure 1.3: Exemple d'AdWords de Google

Le classement des annonces textuelles sur les pages des résultats d'une recherche est l'une des variantes de l'algorithme de Google gardé au plus grand secret afin de préserver la concurrence. Malgré sa complexité, Google a mis en place plusieurs critères laissant aux annonceurs le choix

de bien formuler le contenu de leur AdWords et d'identifier les mots susceptibles d'être saisis par les internautes (voir section « fonctionnement du moteur de recherche »).

La présente étude basée sur la classification des mots-clés se limite uniquement aux données des annonces textuelles.

Mots-clés « Keywords »

Une fois que son annonce est faite, l'annonceur espère qu'elle soit visible sur les pages des résultats lors d'une requête saisie par l'utilisateur le long de sa recherche. Pour cela, il doit, avec précision, définir ces mots qui sont les clefs de son succès. En effet, le choix d'un mot-clé est important étant donné que le système d'enchère développé dans Google associe à chaque mot-clé une valeur d'enchère. Cela donnerait aux annonceurs la motivation pour être mieux placés dans les résultats de recherche.

Rappelons qu'un mot-clé est une expression constituée d'un ou plusieurs mots.

Valeur d'enchère « CPC max »

L'annonceur doit miser sur une valeur pour chacun de ses mots-clés qui sont associés à sa campagne publicitaire. Cette valeur représente le montant maximal ou la valeur d'enchère que l'annonceur accepte de payer pour chaque clic sur son lien sponsorisé obtenu suite à une requête. Dans le moteur de recherche, on ne facture jamais le coût pour un clic supérieur à celui qui est précisé par la valeur d'enchère.

Types de correspondances du mot-clé « Match Type »

Le type de correspondance du mot-clé offert par Google, utilisé par les annonceurs, permet plus de contrôle sur l'apparition du mot-clé ciblé par une campagne publicitaire. La correspondance est un paramètre pour le mot-clé qui définit le degré de correspondance entre le mot-clé et la requête de l'internaute. Ainsi, pour la même requête d'un internaute, l'annonce peut ou pas s'afficher en fonction du type de correspondance.

En voici les principaux types :

Mot-clé exact « Exact »

La requête doit être identique au mot-clé dans sa forme exacte. Ce type de correspondance est souvent utilisé pour des requêtes composées de mots usuels ou souvent saisies.

Expression exacte « Phrase »

La requête doit contenir le mot-clé exact dans sa forme, et complémenté par d'autres mots qui le précèdent ou le suivent.

Par exemple, soit le mot-clé « achat de livre », l'annonce s'affiche sur la requête « achat de livre neuf ». Par contre la requête « achat de beau livre » ne fait pas apparaître l'annonce.

Requête large « Broad »

Cette correspondance est souvent la plus utilisée et offre un large spectre du mot-clé. En effet, l'annonce apparaît dans n'importe quel cas où la requête contient un des mots qui constitue le mot-clé. Il fait apparaitre aussi l'annonce sous forme de synonymes, variantes singulier/pluriel du mot-clé. L'inconvénient de ce type de correspondance est qu'il pourrait facilement épuiser le budget en tombant sur des requêtes non désirées.

C'est pourquoi Google a ajouté une variante aux types de correspondances qui n'est qu'une liste de mots-clés à exclure (ou mots-clés négatifs).

Mot-clé négatif

Les mots-clés négatifs saisis par les internautes ne permettent pas l'affichage des annonces.

Groupe d'annonces « Ad Group »

Dans un groupe d'annonces, l'annonceur trie l'ensemble des mots-clés qu'il veut exécuter dans une campagne en plusieurs groupes et conçoit une ou plusieurs annonces pour chaque groupe spécifique. Les mots-clés dans un groupe sont en général coordonnés par un lien sémantique par rapport aux autres mots-clés. Par ailleurs, un groupe d'annonces est un regroupement d'annonces et de mots-clés qui doivent être en rapport avec le thème du groupe. De plus, cela aide à mieux organiser la gestion des valeurs d'enchère des mots-clés.

Dans certains ouvrages, le terme « créatif » est utilisé au lieu du terme « annonce ». Autrement dit, le terme « créatif » n'est qu'une description de l'annonce.

Campagne publicitaire

La publicité en ligne prend la forme de campagne publicitaire qui consiste à faire la promotion d'une marque, d'un produit ou d'un service. Une campagne est constituée d'un ou plusieurs groupes d'annonces. « Elle est caractérisée par des paramètres qui définissent son budget quotidien, son ciblage géographique et linguistique, ainsi que sa date de début et de fin. Un annonceur peut mener plusieurs campagnes simultanément » (Quinn, 2011).

La structure d'une campagne est de la forme suivante (Figure 1.4)

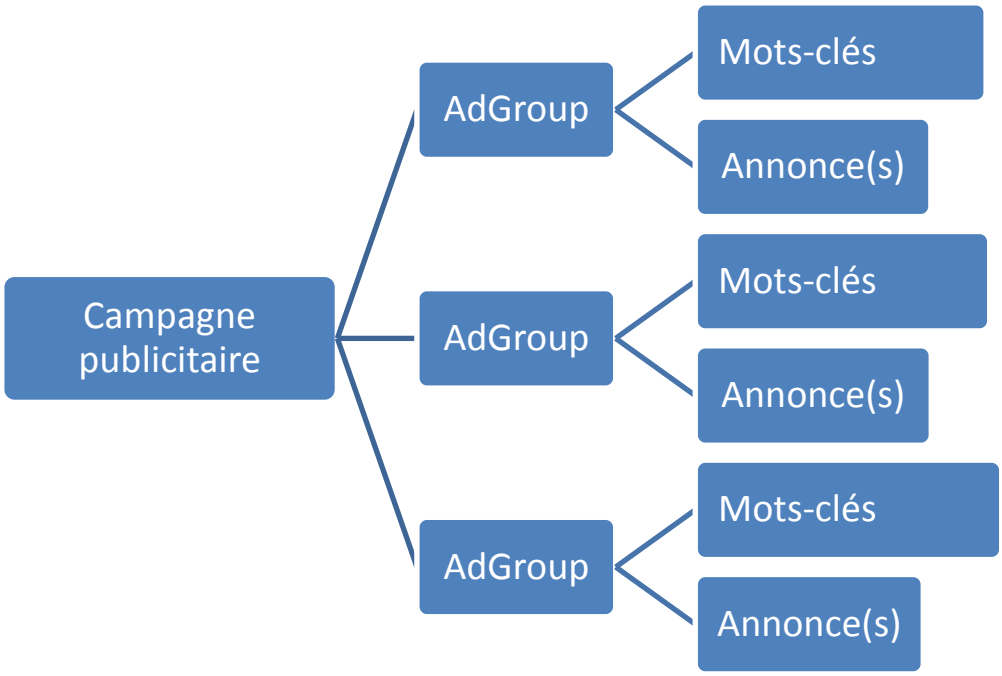


Figure 1.4: Structure d'une campagne publicitaire

1.2 Fonctionnement d'une campagne publicitaire

Le fonctionnement se fait au coût par clic, autant dire que les annonceurs ne paient que lorsque l'on clique sur leurs liens commerciaux. Gérer sa propre campagne publicitaire c'est faire une enchère sur un ensemble de mots-clés et fixer un budget à cette campagne. Les annonces s'affichent jusqu'à épuisement du montant prévu. Ainsi, un rapport quotidien est délivré par le moteur de recherche donnant les statistiques du comportement des mots-clés visés par l'affichage des annonces.

Ces données composées de paramètres sont des indicateurs qui, par la suite, serviront d'atout pour une gestion meilleure de ladite campagne. Parmi ces indicateurs essentiels, on cite :

- Impression

On parle d'impression pour désigner l'affichage d'une publicité sur le Web. L'indicateur impression est le nombre de visualisations de cette annonce apparues dans une journée, reliées au mot-clé visé.

- Clic « Click »

L'indicateur Clic désigne le nombre de clics sur l'annonce associée au mot-clé qui l'a affichée. Ainsi, lorsqu'un internaute est exposé à une publicité et clique sur celle-ci, un clic est comptabilisé.

- Coût « Cost »

À chaque clic, un montant est facturé. Donc, le paramètre coût est la somme des coûts réels facturés dans la même journée.

- Conversion

Toute action jugée importante sur le site Web qu'effectue l'internaute en ayant cliqué sur l'annonce affichée dans le moteur de recherche est une conversion. Cet indicateur est le plus pertinent du moment qu'il nous informe sur la possibilité que l'internaute puisse se convertir en client. Cependant, il est difficile à mesurer vu que l'action ne se fait pas au niveau du moteur de recherche, d'où la possibilité d'absence de la valeur de la conversion dans le rapport du moteur de recherche où la valeur peut avoir lieu mais à une fréquence de temps différente.

- Position

L'indicateur position est la moyenne des positions de l'annonce associée au mot-clé pendant une journée, pondérée par le nombre d'impressions obtenues par chaque position. La valeur de cet indicateur peut ne pas être entière.

Par ailleurs, d'autres indicateurs, très importants, résultent du calcul mathématique en fonction des autres paramètres susmentionnés.
On cite :

- Coût par Clic « CPC »

Il s'agit de la moyenne quotidienne des coûts par nombre de clics. Puisque les données sont fournies sous un format agrégé, alors il n'est pas possible de connaître exactement le montant individuellement payé pour chaque clic.

- Taux de clic « CTR »

Le Taux de clic est le rapport entre le nombre d'impressions d'une annonce et le nombre de clics sur cette annonce, généré par le mot-clé.

$$CTR = \frac{Clic}{Impression}$$

On considère que plus le taux de clic est élevé, plus l'annonce est efficace. Cet indicateur est très utilisé par les gestionnaires de campagnes publicitaires comme étant l'outil de performance et est aussi considéré comme étant un des paramètres de base pour le classement de l'annonce sur le moteur de Google.

Rappelons que certains indicateurs susmentionnés sont sur une base de calcul journalière et il se peut que ces données relatives à ces indicateurs biaisent l'exactitude de l'information. Étant donné qu'on ne connait pas la valeur précise de l'information, ces données sont, par hypothèse, considérées exactes sur le plan de nos analyses.

Classement des annonces sur le moteur de recherche

Les annonces sont classées dans l'ordre des positions et les *n* premières annonces sont alors réparties dans les emplacements de la première page, en respectant cet ordre où *n* est le nombre d'emplacements disponibles sur la page. Si d'autres annonces suivent, celles-là seront réparties sur les pages suivantes en respectant l'ordre des positions (Burriel, 2010).

La position de chaque annonce est calculée à partir de la formule suivante:

$$\textit{Position de l'annonce} = \textit{Enchère maximale} \times \textit{Indice de qualité}$$

L'indice de qualité est un indicateur de pertinence du mot-clé d'une annonce par rapport aux termes d'une recherche. Il est défini et utilisé uniquement par le moteur de recherche de Google. L'indice est calculé sur différents facteurs liés à l'annonce et échelonné de 1 à 10. Parmi ces facteurs, il y a :

- CTR au niveau des mots-clés
- CTR au niveau de la campagne publicitaire
- La concordance des mots-clés avec l'annonce
- La pertinence du mot-clé et la page de destination URL

Avec un indice de qualité élevé l'annonce peut occuper la meilleure position du classement sur un mot-clé sans que la valeur d'enchère soit la plus élevée parmi les autres valeurs d'enchères des annonceurs.

Effet de la position

Le nombre de clics générés par une annonce varie en fonction de la position et cette hypothèse est revérifiée sur une banque de données d'Acquisio par Quinn (Figure 1.5). Ainsi, une position dépend aussi du CPC qu'il faut débourser pour l'atteindre.

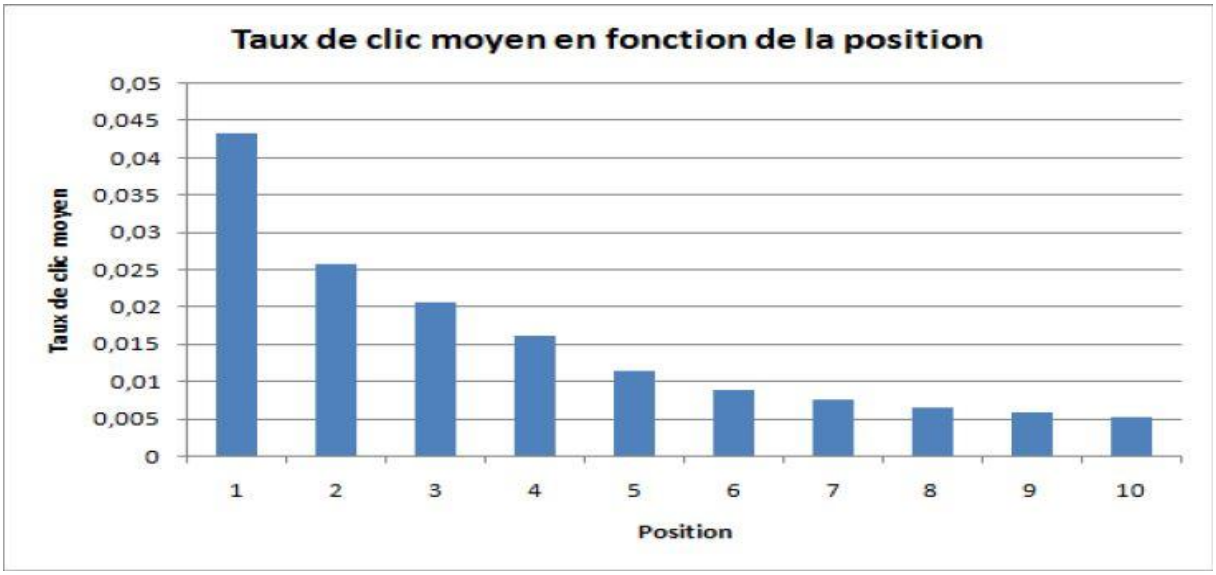


Figure 1.5: CTR en fonction de la position (Quinn, 2011)

Le problème consiste donc à trouver la position optimale pour laquelle une quantité satisfaisante de clics est obtenue à un coût acceptable. Un moyen efficace pour y arriver consiste à utiliser des algorithmes d'optimisation.

L'optimisation repose essentiellement sur les mots-clés qui conviennent pour la campagne publicitaire. Pour cela, une analyse des mots-clés sera réalisée pour l'opération d'optimisation. En effet, une liste de mots-clés potentiels formée par l'analyse de mots clés constitue la base d'une stratégie d'une gestion de campagne publicitaire.

1.3 Description du projet

Les travaux cités ci-dessous ont fait l'objet de recherche au sein de l'entreprise Acquisio et permettent d'améliorer la performance de son logiciel.

Une étude de faisabilité préliminaire d'un système d'optimisation du rendement des campagnes publicitaires (Chan et al, 2008) a tenté de regrouper les mots-clés. Ils se sont basés sur l'hypothèse que les mots-clés n'ont pas tous le même comportement. Par conséquent, le

regroupement des ces mots-clés selon des critères pertinents permettra de définir des relations significatives. Parmi ces critères, on cite :

- Match Type et min CPC.
- Impressions et CTR
- Nombre total impressions et clics

Le procédé n'a pas donné de résultats probants. Une autre tentative de regroupement a été expérimentée mais uniquement sur les mots-clés qui génèrent des conversions. L'expérience a montré qu'il est possible d'établir des relations entre les variables déterminantes pour les mots-clefs engendrant beaucoup de conversions.

Une étude, faite plus tard, réalisée par Quinn (2001), repose sur l'optimisation du budget qui nécessite des estimations des paramètres du modèle (Clic, CPC) de programmation mathématique. En d'autres termes, établir un algorithme qui affecte des positions optimales aux mots-clés de sorte que le nombre total de clics par campagne soit maximisé, en respectant entre autres, la contrainte de budget. Le problème est que, parfois, il n'y a pas beaucoup d'historiques sur certains mots-clés pour effectuer cette estimation. Quinn (2001) a proposé une méthode de génération de courbes génériques pour chaque mot-clé afin d'effectuer une prédiction des paramètres de son modèle mathématique. Le but de ces fonctions est de faire augmenter le potentiel de prédiction de ces paramètres prédictifs Clic et CPC en fonction de la position.

Dans le cadre du présent mémoire, Acquisio a mis à notre disposition une banque de données expérimentales disposant de plus de 27 millions de mots-clés. Ceux-là sont crées par différentes agences à partir de divers campagnes publicitaires. Vu le nombre de mots-clés, une automatisation des analyses est indispensable. En effet, les récents algorithmes de classification permettent d'analyser une quantité importante de données et d'en déduire les critères recherchés.

Avec une meilleure classification des mots-clés on croit pouvoir produire des courbes génériques de meilleure qualité. Pour ce faire, notre travail dans ce mémoire de recherche propose deux principales étapes. Dans la première, nous procédons à une analyse des mots-clés par les méthodes d'exploration des données ou fouille de données, connue sous l'expression data mining, dans le but de déceler des comportements similaires au niveau des mots-clés et de les classifier par la suite. Cette étape consiste à sélectionner, traiter, transformer et appliquer certains

algorithmes de clustering pour interpréter et évaluer les bases de données sources. Les algorithmes appliqués, tels que : k-means, Clara, Pam et Fuzzy C-means (Abonyi, 2007) sont nécessaires pour analyser efficacement ces données. Dans la deuxième partie, nous nous intéressons à l'amélioration des paramètres estimés des courbes génériques sur les différents groupes (clusters) déterminés à la première étape. Nous espérons que ces ajustements sur les fonctions génériques permettront d'effectuer de meilleures prédictions.

CHAPITRE 2 : REVUE DE LITTÉRATURE

Ces dernières années, on trouve un bon nombre d'articles dans le domaine du comportement des mots-clés sur les moteurs de recherches quant aux résultats de recherche naturels (tels que vu dans la section 1.1); c'est à dire les annonces textuelles dont les liens ne sont pas sponsorisés. On y traite des travaux algorithmiques afin d'établir les relations entre les mots-clés. Parmi les techniques utilisées, on cite la méthode de classification automatique hiérarchique qui regroupe les mots-clés par leurs suffixes. C'est ce qui a été exploité par Suneetha, Fatima et Pervez (2011). L'objectif de cette méthode est de répartir les mots-clés en groupes sous différentes contraintes, généralement selon leurs caractéristiques sémantiques, afin que chaque groupe soit le plus homogène possible et qu'il soit le plus distinct comparé aux autres groupes. Ensuite, une relation d'hiérarchie entre les mots est établie par les suffixes, constituant un arbre appelé « dendrogramme ».

Cependant, peu d'ouvrages traitent du comportement des mots-clés sur les moteurs de recherches concernant les liens commandités. Les entreprises qui subventionnent des recherches dans ce domaine ne publient que très rarement leurs résultats à cause de leur forte concurrence. Jusqu'à présent, la plupart des études publiées ont été effectuées en vue d'améliorer le rendement des campagnes publicitaires par des algorithmes d'optimisation (recherche opérationnelle) sur les moteurs de recherche afin de définir l'enchère optimale de chaque mot-clé. On cite entre autres, les articles des auteurs Kitts et Leblanc (2004), Quinn (2011) et Hosanagar et Abhishek (2012).

Dans leur article, Baeza-Yates, Hurtado, Mendoza et Dupret (2005) proposent des modèles pour décrire les modes de comportement des utilisateurs dans les sessions de requête. Ils ont montré que les utilisateurs du Web formulent des requêtes courtes et se contentent de sélectionner quelques pages. On y trouve des illustrations de distribution « clic/position » et la fréquence de la requête qui semble intéressante.

L'article de Rusmevichientong et Williamson (20) suggère aussi un algorithme de sélection de mots-clés pour la recherche des annonces textuelles. L'algorithme se base sur l'historique des performances et l'ordre des préfixes des mots-clés.

Dans un autre article, les auteurs Kitts et al. (2005) avancent également que la relation entre la position et les clics et la relation entre l'enchère et la position suivent une fonction exponentielle paramétrée et qu'elles fournissent des prédictions analytiques sur les ventes.

Les travaux de Chan et al. (2008), déjà susmentionnés, ont révélé l'existence de critères pertinents pour créer des classes de mots-clés. Ils ont modélisé le comportement entre la position moyenne, les impressions et les clics pour les mots-clés engendrant beaucoup de conversions. Le nombre de mots-clés est limité, voire trop peu élevé dans ces cas. Des hypothèses et suggestions très intéressantes ont été formulées.

Par contre, l'analyse des mots-clés par les méthodes de classification demeurent confidentielle (Duijndam, 2012, chapitre 3-4).

Le développement des outils aux services des moteurs de recherche ne cesse pas de s'étendre. Depuis mai 2011, un nouvel outil a été lancé sur internet via Google « Google Correlate » utilisant les méthodes d'exploitation et d'analyse des données issues du moteur de recherche et permettant d'établir des corrélations qui correspondent aux tendances des mots ou des expressions sur une période donnée (Mohebbi, 2011). De plus, il y a l'outil du Web analytique qui s'appuie sur les méthodes de data mining et fournit des informations sur les internautes d'un site Web. Le but en est de générer une catégorisation des visiteurs de sites pour mieux en cibler les consommateurs et optimiser les performances de la campagne publicitaire.

L'exploration de données, souvent utilisée en terme « data mining », est l'ensemble des techniques d'analyse et de compréhension des données. De telles techniques sont, de nos jours, très utilisées dans divers domaines.

Même dans un environnement très confidentiel, cela n'exclut pas que l'appel aux méthodes de data mining soit pris en considération vu le nombre de données multidimensionnelles où il est très difficile de définir la distribution qui les régit.

L'objectif de ce chapitre est de donner une présentation globale du domaine de data mining ainsi que des méthodes et algorithmes qui en découlent, afin de permettre une meilleure compréhension de ce travail.

2.1 Data mining

Tufféry (2007) définit le data mining comme un ensemble d'algorithmes et de méthodes destinés à l'exploration et à l'analyse de grandes bases de données informatiques dans le but de détecter des règles, des tendances inconnues ou des structures particulières. Il mentionne également l'importance du calcul et de la représentation visuelle. Fayyad, Piatetsky-Shiparo et Smyth (1996), à leur tour, définissent que le data mining est une étape dans le processus d'extraction des connaissances à partir des données (ECD) qui consiste à appliquer l'analyse des données et des algorithmes de découverte produisant un recensement particulier de modèles sur les données.

Lors d'une conférence, l'auteur Saad (2012) expose le data mining sous différentes perceptions algébriques, analytiques et statistiques. Plusieurs graphes sont exposés, donnant un point de vue sur la complexité des modèles, et par différentes techniques de projection utilisées sous divers angles ou par décomposition matricielle, un certain cas de similarité entre les données est repéré. Des formules mathématiques sont éclaircies afin de les utiliser dans le matériel informatique.

Parmi les principales techniques de data mining, on distingue les techniques descriptives et prédictives.

Tufféry (2007) souligne que les techniques descriptives, que l'on appelle également « techniques non supervisées », visent à mettre en évidence des informations présentes mais cachées par le volume des données, à réduire, à résumer et à synthétiser les données. De même, il n'y a pas de variables cibles à prédire. En outre, les techniques prédictives, appelées aussi « techniques supervisées », visent à extrapoler de nouvelles informations à partir des informations présentes, d'où la présence d'un modèle à variables cibles à prédire.

Il est préférable d'établir un vocabulaire de certains termes techniques que nous jugeons essentiel vu qu'il existe une différence de syntaxe ou de compréhension chez certains auteurs en passant de l'anglais vers le français.

2.2 Vocabulaire

Du point de vue vocabulaire, la communauté scientifique francophone utilise différents termes pour désigner le mot anglais « clustering » qui est communément employé.

Agard (communication personnelle, janvier 2012) dans son cours, évoque la difficulté de trouver le bon mot en français : on parle de classification, partition, segmentation ou regroupement des données.

Le terme « classification » en anglais fait référence à l'affectation d'un individu, par des méthodes de probabilités (ex : Bayes), à une classe existant a priori. Cela se traduit en français par le terme « classement ».

La distinction entre l'apprentissage supervisé et non supervisé réside dans l'étiquetage des données selon leur catégorie : cela veut dire la connaissance ou non des classes. Dans les méthodes où l'on utilise les données étiquetées par leur appartenance à une catégorie, on parle d'apprentissage « supervisé »: citons un exemple bien connu du jeu de données de fleurs d'iris, ou chaque fleur est étiquetée par son appartenance à une espèce bien définie. Et dans celles où l'on utilise les données non étiquetées, sans leurs catégories, on parle d'apprentissage « non supervisées ». Les «classificateurs» sont des méthodes supervisées et les «clustering» des méthodes non supervisées.

Le terme anglais « cluster » fait référence à une classe, à un groupe ou encore est appelé « grappe », obtenu par les méthodes de classification.

Les termes «données», «enregistrements de données», «vecteurs de données» et «ensemble de données» désignent un ensemble d'éléments que l'utilisateur souhaite extraire.

2.3 Méthodes de fouille des données

Au cours de ces dernières années, de nombreux travaux proposent différentes méthodes pour déceler certains comportements dans une base de données telles que le « clustering ». Le problème de clustering a été formulé de différentes manières en statistique, la reconnaissance de formes, les bases de données, l'optimisation, etc.

Dans le domaine de la fouille de données, le clustering est souvent utilisé comme outil de visualisation d'un ensemble de données dans le but d'y découvrir une possibilité de structure de groupe latente. Il est aussi utilisé pour compresser un ensemble de données en le remplaçant par un ensemble de prototypes.

2.3.1 Techniques non supervisées

Certains auteurs définissent cette méthode comme étant une source de découverte fortuite. Et vu que l'on ne dispose que de données non étiquetées et que le nombre de classes et leur nature n'ont pas été prédéterminés, alors la structure plus ou moins cachée des données doit être découverte par les algorithmes eux-mêmes (Wikipédia, 2013).

Dans une classification non supervisée, de nombreux ouvrages mentionnent qu'il revient à l'utilisateur de spécifier le problème à résoudre et le choix du nombre de classes. Il existe des algorithmes de classification permettant de choisir le nombre de classes en se basant sur des indices de mesure de qualités du clustering.

Duda, Hart et Stork (2001) fournissent de meilleures recommandations dans leur ouvrage concernant les premières étapes d'une étude d'un apprentissage non supervisé. Ils conseillent fortement d'avoir un aperçu de la nature ou de la structure des données. Ils commencent par l'hypothèse très restrictive que les formes fonctionnelles des densités de probabilité sous-jacentes sont connues et que la solution sera conduite à diverses tentatives de reformulation du problème comme étant un problème de partitionnement des données en sous-groupes. Duda et al (2001, chapitre 10) mentionnent aussi que certaines des procédures de regroupement n'ont pas de propriétés théoriques connues et elles sont toujours parmi les outils les plus utiles.

Les données sont ciblées selon leurs attributs pour les classer en groupe homogènes. La similarité est généralement calculée selon une fonction de distance entre paires de données. C'est ensuite qu'on doit associer ou déduire du sens pour chaque groupe et pour les motifs d'apparition de groupes dans leur espace. Notons que la plupart des algorithmes de classification ont pour point de départ une mesure de distances, ou dissemblances, entre les objets. Dans les ouvrages de Zaki et Meira (2013) et de Gan, Ma et Wu (2007), les auteurs proposent plusieurs méthodes de clustering qui pourraient être adaptées.

2.3.2 Techniques supervisées

L'apprentissage supervisé est le thème le plus traité dans plusieurs ouvrages, notamment celui de Hastie, Tibshirani et Friedman (2009), où les classes sont prédéterminées et les exemples sont connus et étiquetés. Les techniques supervisées visent à établir un classement selon un modèle

considéré comme étant la distribution d'un ensemble de données en groupes ou en classes, en fonction de certaines relations communes ou des similitudes. Compte tenu des *n* différentes classes, un algorithme classificateur construit un modèle qui prédit l'appartenance de chaque enregistrement non étiqueté de données à une classe avec une grande de probabilité.

Quelques algorithmes de classification

Plusieurs algorithmes ont été proposés et sont disponibles dans la littérature. La méthode k-means (Hartigan et Wong, 1979) est considérée comme l'une des meilleurs en termes de temps et de résultats. Dans l'ouvrage « data mining and knowledge discovery series », les auteurs comptent l'algorithme de k-means parmi les dix meilleurs des algorithmes de data mining (Wu & Kumar, 2009, chapitre 2).

Bradley, Fayyad et Reina (1998) mentionnent que le clustering est une étape cruciale de l'exploration de données et que l'exécution des outils d'exploration des bases de données massives est essentielle. Ils ont mis l'accent sur les approches k-means malgré que de telles techniques soient basées sur des notions de métriques de distance et qu'elles ne permettent pas à un ensemble de données d'avoir la possibilité d'appartenir à différents groupes. Dans cet article, les auteurs se concentrent sur la tâche de mise à l'échelle de la technique clustering probabiliste appropriée qui est l'algorithme E-M « Expectation-Maximisation » (Dempster, Laird et Rubin, 1977). Elle a des propriétés qui ne nécessitent pas la spécification des mesures de distance et admet les attributs non continus.

L'algorithme k-means tente de minimiser la somme des carrés des distances euclidiennes entre les enregistrements de données dans un groupe et le vecteur moyen de ce groupe. Il suppose que tous les groupes sont représentés par des distributions gaussiennes sphériques. En d'autres termes, il ne détecte que les formes convexes. Et puisque l'algorithme k-means fait appel à la métrique euclidienne, il ne se généralise pas au problème de regroupement des données discrètes ou catégorielles (sauf pour une éventuelle transformation de données). L'algorithme k-means suppose également que chaque enregistrement de données appartient exactement à un groupe.

Bouveyron (2012) propose une modélisation et une classification des données à grande dimension. Il présente, entre autres, son algorithme « HDDC » (High Dimensional Data Clustering) dans le cadre d'une classification non supervisée de modèles gaussiens, suivi de démonstrations mathématiques des propositions. Partovi Nia et Davison (2012) proposent aussi l'algorithme « bclust » (Bayesian clustering), qu'on le retrouve dans R et qui mit en œuvre une approche bayésienne de regroupement des données continues à grande dimension, avec a priori les paramètres du modèle. Quelques applications de l'algorithme sont présentées aux divers domaines sur des données numériques ou non numériques.

La classification par l'approche bayésienne trouve une solution localement optimale pour une fonction objective appelée log-vraisemblance. Comme la surface de la fonction de log-vraisemblance est non convexe, des méthodes « MCMC » (Markov Chain Monte Carlo) ou méthodes de simulation bayésienne ont été proposées (Matusevich, Ordonez et Baladandayuthapani, 2013) ainsi que l'algorithme « mcclust » (MCMC clustering) de Fritsch[1] (2009) implémenté dans R.

Les auteurs Wang, Qiu et Zamar (2007) proposent l'algorithme « Clues » de classification non paramétrique basée sur le rétrécissement locale pour estimer le nombre de groupes et le partage des points de données sans aucun paramètre d'entrée, sauf un critère de convergence. Chaque point de données est transformé de telle sorte qu'il se déplace d'une distance spécifique vers le centre du « cluster ». La direction et la taille associées à chaque mouvement sont déterminées par la valeur médiane de ses *K* plus proches voisins. Ce processus est répété jusqu'à ce qu'un critère de convergence prédéfini soit satisfait. La valeur optimale du nombre de voisins est déterminée par l'optimisation de l'indice de mesure de qualité des groupes générés par l'algorithme.

Kaufman et Rousseeuw (1990, chapitre 2) décrivent l'algorithme « PAM » comme une variante de l'algorithme k-means. Contrairement à l'approche de k-means, PAM accepte la matrice de dissemblance. Il est plus robuste aux valeurs extrêmes car il réduit au minimum une somme des différences au lieu d'une somme de distances et il permet de choisir le nombre de clusters.

[1] http://cran.r-project.org/web/packages/mcclust/mcclust.pdf

Scepi (2010) cite aussi que l'algorithme PAM est une optimisation itérative qui combine la relocalisation de points entre les clusters en perspective avec une nouvelle nomination des points.

Le problème de décider du nombre de classes ainsi que l'évaluation des résultats du clustering a fait l'objet de plusieurs travaux de recherche, dont celui de Gath et Geva (1989). Il y a d'autres pratiques de mesures de qualité des clusters telles que la statistique Gap proposés par Tibshirani, Walther et Hastie (2000) ou encore l'indice « Silhouette » présentés par Wang et al. (2007). C'est avec cet indice que ces derniers ont obtenu les meilleurs résultats lors de l'application des différentes mesures de qualité sur leurs exemples.

Vu la multitude des algorithmes de classification, Benzécri (1980) indique qu'on ne doit pas oublier la multitude des traitements préliminaires des données qui sont généralement décisifs pour la qualité des résultats.

Finalement, la méthode sur la performance du modèle de classement, telle que mentionnée dans l'ouvrage de Bradley et Tibshirani (1993), améliore la performance du modèle prédictif et informe sur sa stabilité. Son principe découpe les données en deux ensembles tirés aléatoirement : une partie constitue un ensemble d'apprentissage de données afin d'estimer le modèle et l'autre partie est utilisée pour tester le comportement de ce modèle.

Dans cette section de littérature, on a procédé de façon globale à la synthèse du processus du data mining ainsi qu'aux différentes approches de clustering. Dans le prochain chapitre, on exposera le déroulement de notre étude que l'on fera suivre des techniques de traitement de données et des méthodes de classification faisant l'objet de notre travail.

CHAPITRE 3 : MÉTHODOLOGIE

Dans le chapitre de la littérature, les travaux publiés étaient généralement concentrés sur les stratégies optimales des rendements des campagnes publicitaires qui paient pour les clics de publicité de mots-clés. Ces études visaient à analyser les données de résultats des campagnes afin d'en extraire des relations mathématiques, telles que les relations (Clic, position) et (CPC, position) qui sont souvent utilisées dans la littérature. Quinn (2011) a procédé à l'élaboration des fonctions génériques, selon les relations déjà mentionnées, permettant de prédire les positions souhaitées des mots-clés sur les moteurs de recherche. Ces fonctions semblent s'éloigner considérablement des valeurs observées dans certains cas. Pour améliorer ces fonctions, nous avons choisi d'exploiter au grand nombre les données relatives aux mots-clés par des approches et méthodes en data mining.

Dans ce chapitre nous allons décrire la démarche scientifique proposée selon les étapes successives de notre expérimentation, à partir des références vues en littérature.

Un travail de conception du modèle y fera également l'objet de notre étude aux moyens des caractéristiques des mots-clés recensées par Google. D'autres caractéristiques sont déterminées par les commentaires faits suite aux questions posées aux gestionnaires de campagnes (section transformation, page 33).

3.1 Objectifs

Tel que nous l'avons déjà mentionné, il s'agit de décrire les méthodes utilisées dans le but d'améliorer les résultats de l'étude antérieure de Quinn par l'approche du data mining avec les techniques de classifications. Pour cela, certains objectifs spécifiques devront être atteints.

Objectifs spécifiques

- Identifier un nombre de classes (groupes) dans une campagne.
- Évaluer la possibilité de généraliser ce nombre de classes pour toutes les campagnes.

Afin de résoudre ce problème d'uniformité du comportement des campagnes, on opte pour une fusion de différentes campagnes, faisant en sorte d'augmenter seulement le volume des mots-clés. Elle sera envisageable pour des éventuels tests. Dans un premier temps, on propose de vérifier, pour chaque campagne, s'il existe une ou plusieurs séparations de mots-clés selon des critères définis a posteriori. Ensuite, une approche générale où en reproduisant les même expériences effectuées cette fois-ci sur plusieurs modèles de campagnes, tirées aléatoirement ou prédéterminées, dans le but d'obtenir plus d'informations au niveau des données relationnelles.

Une fois le nombre de classes choisi, on procède comme ceci :

- Numéroter les mots-clés d'une campagne selon leur classe trouvée.
- Ajuster les courbes génériques des mots-clés selon leur appartenance au groupe.

Cependant, les analyses qui seront effectuées permettront de poser quelques hypothèses, dont la teneur suivra dans la section suivante, comme base de connaissance pour le développement de notre étude.

3.2 Hypothèses

Dans cette section, on formulera les hypothèses utilisées à partir des travaux effectués par les auteurs, tel que vu au chapitre 2, afin de procéder à notre travail. Ces hypothèses sont les suivantes :

- les données agrégées par Google, sur une base journalière, sont supposées exactes sur le plan de nos analyses. On sera contraint d'accepter les données estimées comme unique source d'information.
- les mots-clés n'ont pas tous le même comportement (Chan et al., 2008).
- les taux de décroissance relatifs du nombre de clics et du coût par clic sont relativement constants d'un mot-clé à l'autre (Quinn, 2011).
- On ne tient compte que d'un nombre limité de positions (Baeza-Yates et al., 2005; Quinn, 2011)
- Le nombre de mots saisis dans une requête est significatif (Baeza-Yates et al., 2005).

– L’échantillon de quelques campagnes publicitaires ciblées par les mots-clés est représentatif.

3.3 Étapes de la méthodologie

La partie méthodologie présente la démarche scientifique à suivre afin d’atteindre l’objectif de ce mémoire, telle qu’illustrée à la Figure 3.1. Des méthodes et des outils d’analyse sont empruntés des ouvrages et des articles susmentionnés.

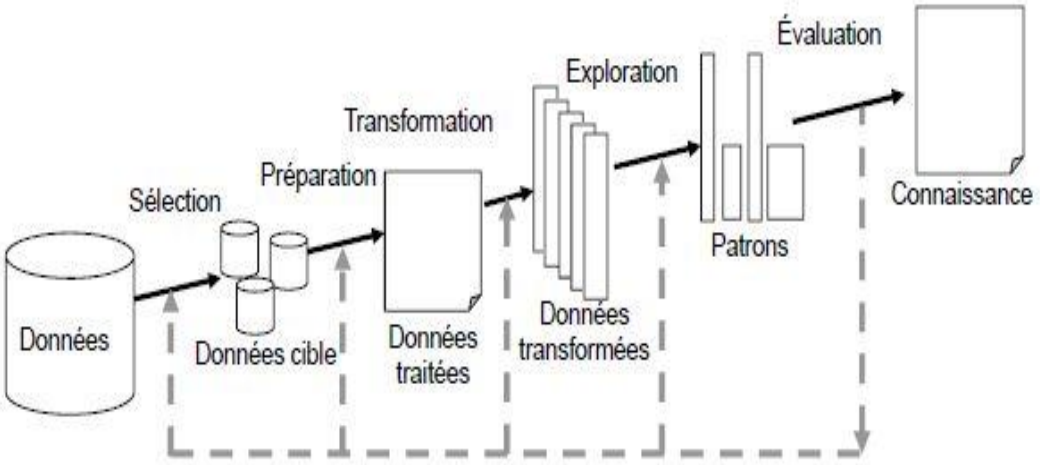


Figure 3.1: Processus d'un ECD. (Fayyad & al., 1996)

La première étape de la méthodologie proposée est l’exploration des données à partir de sources de tables relationnelles gérées par le système de gestion « Oracle MySQL ». Il s’agit de définir les variables des données qui font l’objet de notre modèle de classification.

3.3.1 Présentation des données

Dans cette section, on fournit une description des caractéristiques d’un mot-clé en s’appuyant sur différents modèles à analyser par les méthodes de data mining. On s’y restreint aux données non-confidentielles.

Cela est illustré dans le tableau suivant où chaque mot-clé correspond à un enregistrement de données. Voici les informations que contient un mot-clé :

Tableau 3.1: Description des caractéristiques d'un mot-clé

Caractéristiques	Description	Type
Identifiant	Entité du mot-clé	Caractère
Texte	Expression du mot-clé	Texte
Ad Group	Groupe d'annonces où le mot-clé est affecté	Caractère
Créative	Annonce textuelle	Caractère
Match type	Correspondance du mot-clé	Discrète
Impression	Nombre d'apparitions par jour de l'annonce sur le moteur de recherche activé par le mot-clé	Numérique
Position	Position moyenne de l'annonce sur le moteur de recherche activée par le mot-clé	Numérique
Clic	Nombre de clics par jour de l'annonce apparue sur le moteur de recherche par le biais du mot-clé	Numérique
Coût	Coût par clic sur l'annonce activée par le mot-clé pendant une journée	Numérique
Conversion	Nombre de conversions dans la journée.	Numérique
Temps	Date d'apparition de l'annonce activée par le mot-clé	Caractère

Ces caractéristiques sont les variables principales et leur sélection joue un rôle important sur les résultats des algorithmes de classification.

Les informations reliées au mot-clé sont structurées en deux étapes.

1er étape : **Conception du mot-clé**

À partir d'une formulation textuelle à laquelle on fait correspondre l'attribut match type, un code y sera attribué et il sera l'unique identifiant du mot-clé dans une campagne donnée. Ainsi, pour un même mot-clé de différents Match types, on aura deux identifiants (cela ne pose pas de

problème du moment que les espaces des identifiants sont automatiquement différents). Ce mot-clé est mis dans une liste parmi tant d'autres (voir structure d'une campagne publicitaire, Figure 3.3). Il sera associé aux différents créatifs dans le but d'augmenter les chances de sélection de l'annonce au niveau de l'algorithme de Google. Ces deux ensembles, mots-clés et créatives, sont regroupés conjointement dans un groupe d'annonce « Ad Group ». Par la suite, on attribue à chaque mot-clé un montant d'enchère (CPCmax) où l'on espère faire apparaître l'annonce sur le moteur de recherche en position souhaitée.

2ème étape : **Comportement du mot-clé**

Une fois la campagne publicitaire lancée sur le moteur de recherche, les données statistiques, déjà vues dans la section 1.1 relatives aux mots-clés, seront transmises quotidiennement au niveau du serveur d'Acquisio.

À partir de ces étapes, une illustration (Figure 3.2) de ce processus de relations entre les caractéristiques par rapport au mot-clé favorisera mieux la compréhension du comportement du mot-clé.

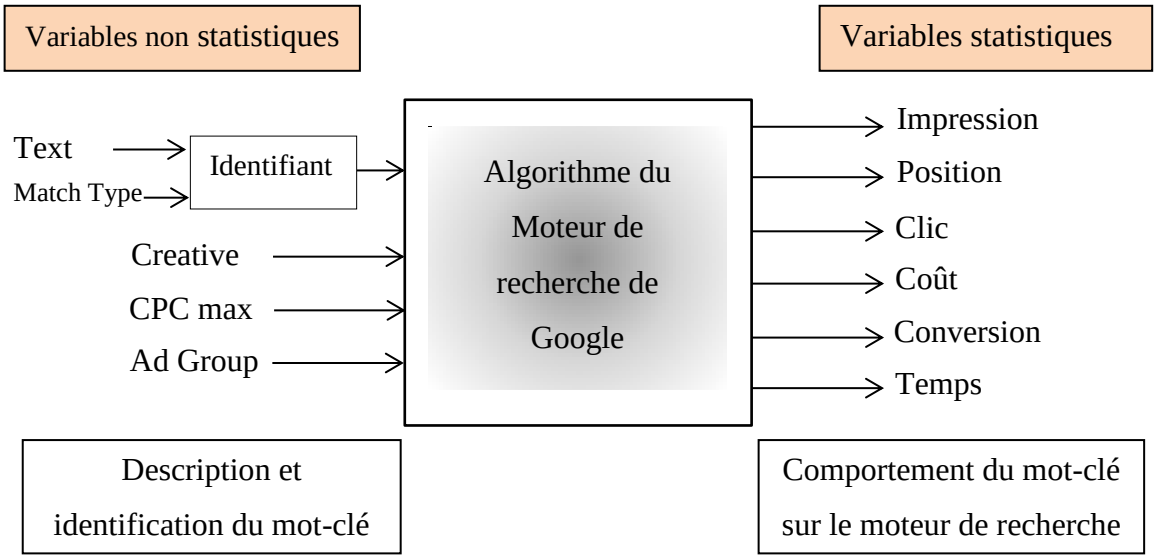


Figure 3.2: Modèle de données d'un mot-clé

Ce processus de description est important, car il permet d'analyser et d'encadrer les données afin de mieux cibler les variables que nous estimerons favorables dans l'étude des techniques de classification non supervisée. Malheureusement les variables CPCmax et Indice de qualité ne sont pas dans notre banque de données.

3.3.2 Extraction des données

À partir des bases de données fournies par Acquisio, structurées de la façon du modèle représenté par la Figure 3.3, une requête SQL permet d'extraire les caractéristiques susmentionnées se rapportant aux activités quotidiennes des mots-clés d'une campagne publicitaire reliée à une agence définie, par le moteur de recherche Google. Les données d'une campagne publicitaire sont formées de n individus (code de mots-clés) représentées sur p variables (caractéristiques suscitées) numériques ou textes.

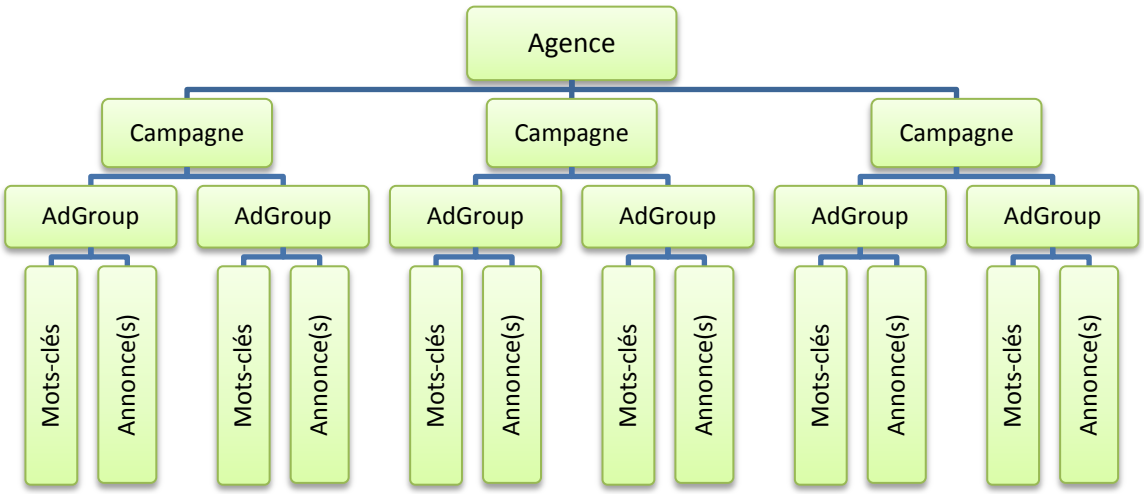


Figure 3.3: Organisation des campagnes à partir d'une agence publicitaire

Ci-dessous, un exemple de données de comportement quotidien de mots-clés d'une campagne publicitaire reliée à une agence donnée.

Tableau 3.2: Données quotidiennes relatives au comportement des mots-clés sur le moteur de recherche

	ID_MOTCLÉ	TEXT	ADGROUPID	CREATIVEID	MATCHTYPE	IMPRESSION	POSITION	CLIC	COÛT	CONVERSION	ID_TEMPS
1	145104	xxx1	29334	13774	Broad	10	8,8	0	0	0	4026
2	145104	xxx1	29334	13784	Broad	7	10	0	0	0	4026
3	145464	yyyy1	29354	13854	Broad	1	9	0	0	0	4026
4	145604	yyyy2	29364	13884	Broad	1	9	0	0	0	4026
5	146874	zzz1	29394	14004	Broad	1	15	0	0	0	4026
6	146874	zzz1	29394	14014	Broad	3	10,33	0	0	0	4026
7	146884	zzz1	29394	14014	Exact	1	15	0	0	0	4026
8	149304	aaa1	29424	14114	Broad	7	11,43	0	0	0	4026
9	169144	nnnaa2	30004	16314	Broad	1	2	2	1,28	0	4026
10	…	…	…	…	…	…	…	…	…	…	…
11	…	…	…	…	…	…	…	…	…	…	…
116	…	…	…	…	…	…	…	…	…	…	…
117	193764	aaaa2	28654	18924	Phrase	1	2	0	0	0	4026
118	193774	aaaa2	28654	18924	Exact	1	9	0	0	0	4026
119	196354	bbb1	28764	19364	Broad	2	10,5	0	0	0	4026
120	197054	bbbaa1	28804	19494	Broad	3	8	0	0	0	4026
121	199364	ccbba	28854	19754	Broad	1	6	0	0	0	4026
122	199484	cccbba	28864	19794	Broad	2	8	0	0	0	4026
123	199534	ggrf	28864	19794	Broad	2	9,5	0	0	0	4026
124	201444	cferr	28944	20104	Broad	2	33	0	0	0	4026
125	201554	eeeff	28954	20134	Broad	1	7	0	0	0	4026
126	201724	eeeffdd	28954	20144	Broad	1	11	0	0	0	4026
127	211394	eeeffs	29184	21074	Broad	29	9,34	1	0,69	0	4026

Les variables sont en colonnes et les observations ou les individus (identificateur de mots-clés) sont en ligne.

Pour des raisons de confidentialité, les codes des identificateurs des variables ainsi que les textes des mots-clés ont été changés, mais gardant la même forme des valeurs initiales. De plus, les clients associés aux agences ne sont pas mentionnés.

Dans notre banque expérimentale de données d'AdWords, on a recueilli les statistiques suivantes :

Nombre d'agences : 81

Nombre de campagnes : 14 893

Nombre de mots-clés : 28 241 431

Afin de mener à bien notre analyse, un échantillonnage de données est prérequis et fondamental. D'après Tufféry (2006), « un bon échantillonnage est toujours délicat à réaliser et nécessite une bonne connaissance de la population ».

Échantillonnage

Tel que déjà vu, nous nous intéressons aux comportements de mots-clés dans une campagne publicitaire. Cela révèle que notre population visée est représentée par les mots-clés. Ainsi, l'échantillonnage sera une partie de ces mots-clés. Or, d'après l'organisation des campagnes publicitaires à partir des agences (Figure 3.3), les mots-clés sont au dernier niveau de la formation. De plus, chaque mot-clé a son propre identifiant dans sa campagne. Cela revient à dire que pour un même mot-clé de même définition, il y a différents identifiants dans plusieurs campagnes publicitaires.

Exemple :

Tableau 3.3: Exemple de codage de mots-clés dans divers campagnes

Caractéristique	Campagne	Code mot-clé	Texte	Créative	Match type
Mot-clé 1	A	0011	achat livre	ABC	« Broad »
Mot-clé 2	B	0211	achat livre	ABC	« Broad »

Ces renseignements nous amènent à choisir un échantillon de campagnes parmi ceux fournis par les diverses agences pour lesquelles nous disposons beaucoup de données concernant l'évolution des mots-clés. Nous choisissons en particulier des campagnes qui engendrent un nombre important de clics.

Jusqu'à présent, on n'a présenté que des données sur une échelle temporelle d'une journée. Mais dans la réalité, une campagne peut prendre plusieurs jours, voire des mois, ce qui explique l'étude des comportements des mots-clés. Considérons qu'une observation d'un mot-clé est un enregistrement de données statistiques d'une journée, transmis par Google, la structure globale des données relationnelles d'une campagne publicitaire sur une période de plusieurs jours sera illustrée comme suite :

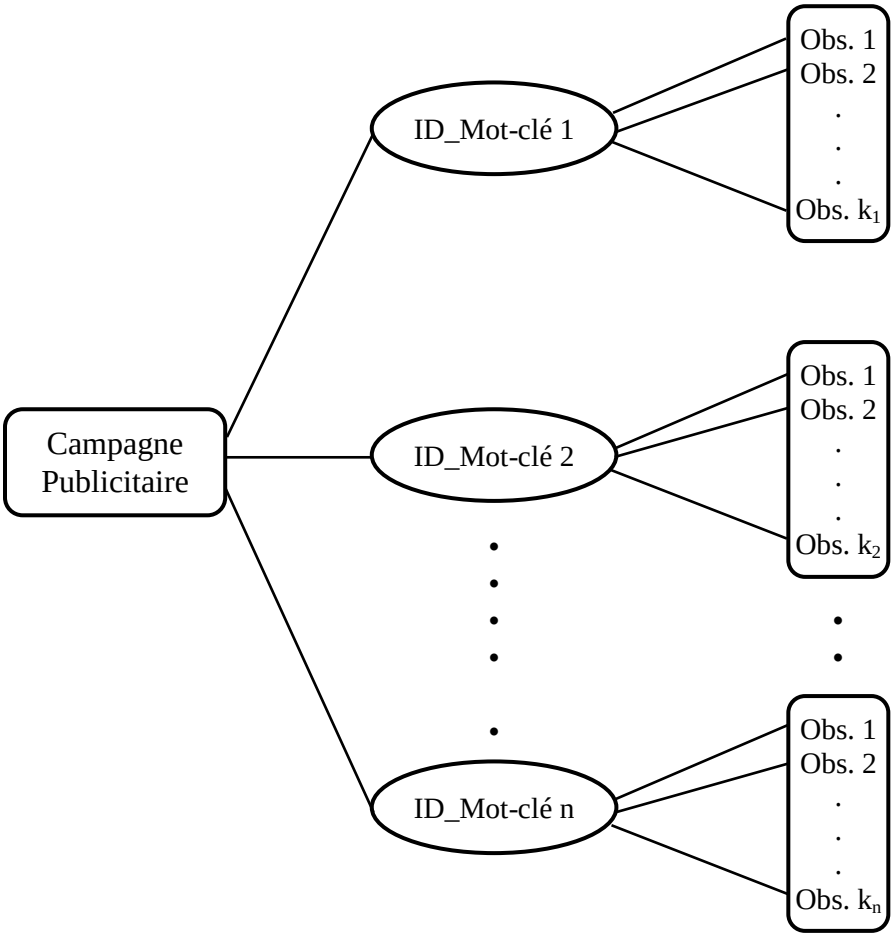


Figure 3.4: Structure de données relationnelles d'une campagne sur une période de plusieurs jours

Étant donné qu'il y a un volume important de données d'observations de mots-clés dans des campagnes publicitaires, la réduction de données peut dissimuler certaines informations. Pour cela, on s'assure, dans un premier temps, que les mots-clés dans une campagne soient représentatifs et de taille non réduite[2] pour visualiser l'ensemble des comportements des mots-clés dans un espace multidimensionnelle. Cependant, un problème de redondance de données s'impose. On y reviendra dans la section 3.3.3.

Dans un second temps, il serait intéressant d'étudier une autre échelle (semaine, mois) en fonction des besoins de l'analyse.

Comme le nombre de dimensions est supérieur à trois, la visualisation simultanée de l'ensemble de données demeure un problème. Des opérations de manipulation de données visant à présenter des vues, telles que les projections, les rotations, doivent être appliquées afin de rendre possible la visualisation.

Pour constituer notre échantillon à partir de ces informations, on procèdera à la manière suivante :

- un échantillon de campagnes est extrait aléatoirement dans différentes agences afin de s'assurer qu'il soit représentatif.
- un échantillon est extrait selon la compatibilité de campagnes publicitaires similaires. (Exemple : des campagnes de ventes d'automobiles de marques différentes). Cette collecte est faite en évitant l'introduction d'information biaisée.
- un échantillon de campagnes de natures totalement divergentes (non similaires) afin de vérifier la compatibilité des résultats de données de mots-clés des campagnes.

La taille de notre échantillon a été établie à partir de 200 campagnes réparties sur 30 agences dont le nombre d'observations de mots-clés dépasse les 15 000.

L'échantillon sur lequel porte notre analyse est très restreint et l'ajout d'un certain nombre d'échantillons additionnels permettra de confirmer certains résultats.

---

[2] Après avoir effectué les opérations de pré-traitement

3.3.3 Prétraitement des données

Cette étape est primordiale : elle nous permet d'identifier et de sélectionner les variables à inclure dans notre analyse de classification non supervisée. Pour cela, certaines tâches devront être respectées.

Nettoyage : valeurs manquantes, valeurs aberrantes

L'existence d'erreurs au niveau de la saisie de données et la nécessité de ressaisir les données sont écartées, car toute la saisie de données est automatisée; l'intervention humaine dans le processus est éliminée. Cela explique que les statistiques reçues par Google reflètent l'exactitude du comportement du mot-clé sur le moteur de recherche.

Sachant que les conversions sont rares (Quinn, 2011), pour plusieurs mots-clés ces valeurs sont presque nulles, alors la variable en question sera éliminée.

Transformation et ajout :

Généralement le mélange de variables qualitatives et quantitatives pose quelques difficultés. Mais dans notre cas, la variable Match type sera transformée en forme binaire.

Tel qu'expliqué à la section 1.2, les variables CTR et CPC seront ajoutées avec les données statistiques qui sont des indicateurs très pertinent vis-à-vis au gestionnaire de campagne.

D'après notre sondage, l'ajout de la variable « nombre de mots dans un mot-clé » sera significatif, ce qui confirme l'hypothèse de Baeza-Yates et al., (2005).

Réduction :

Les redondances des variables mesurant le même phénomène ou une variable dont la valeur s'obtient à partir de combinaison mathématique de valeurs des autres variables seront évitées (problème de colinéarité). Ce n'est cependant pas le cas pour les deux variables CTR et CPC, respectivement en combinaison avec les variables (Impression et Clic) et (Coût, Clic). Celles-là feront l'objet d'analyse de classification non supervisée très attentive en présence des deux variables qui les constituent.

Le nombre de positions, tel qu'il est mentionné dans les hypothèses, sera limité à 12 (voir section 3.2). Il s'agit du nombre d'annonces affichées généralement sur les premières pages de résultats du moteur de Google. Il y aura certes quelques valeurs extrêmes que nous jugerons peu influençables dans notre analyse car la plupart des mots-clés qui sont aux positions 13 et plus ont des caractéristiques nulles et sont faibles en portion ou en quantité. De plus, l'objectif des agences est de faire apparaître leurs annonces à la première page. Illustrons un exemple d'un cas concret d'une campagne publicitaire extrait de notre base de données par un graphique montrant la répartition des valeurs de la variable position (Figure 3.5). On constate que la justification de l'hypothèse de la limite du nombre de positions est nettement vérifiée.

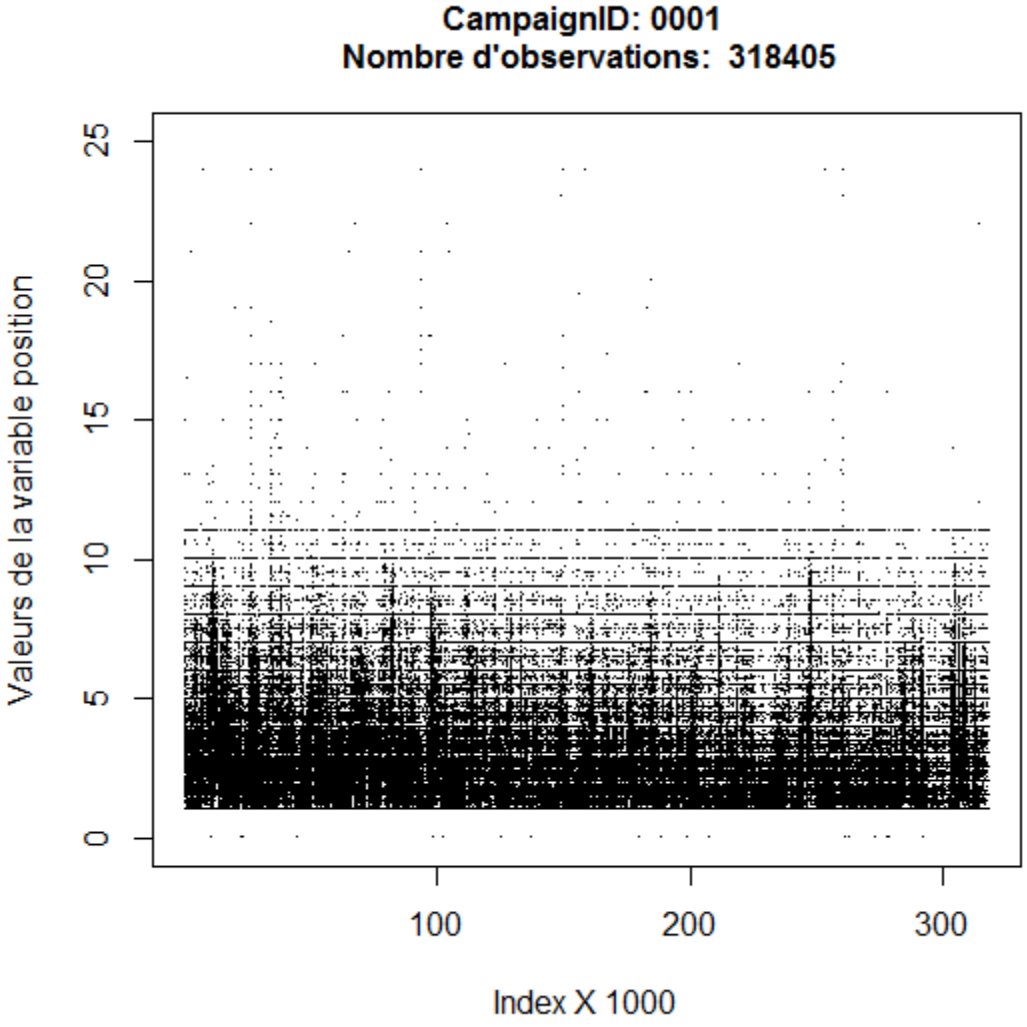


Figure 3.5: Exemple de représentation graphique de la variable position

Outils d'analyses

Les principaux outils utilisés pour cette étude font partie des fonctionnalités du logiciel R. Rappelons que R est un langage de programmation mathématique et statistique, créé par Ross Ihaka et Robert Gentleman[3] dans les années 90, utilisé surtout pour manipuler, traiter et analyser les données. De nombreuses techniques statistiques y sont implémentées. En plus de ses fonctionnalités de base, R dispose de librairies de fonctions très développées à travers ses modules externes, appelés « packages », qui sont tout à fait adéquates pour l'affichage de graphiques évolués et le calcul des méthodes avancées qui sont très flexibles et utiles telles que le clustering.

3.3.3.1 Gestion des redondances des mots-clés

Comme nous l'avons déjà mentionné à la section d'échantillonnage, tous les mots-clés d'une campagne publicitaire seront étudiés simultanément sur une échelle temporelle journalière. Une répétition des identifiants sera établie dans l'espace des individus et peut affecter la classification. En effet, un même identifiant peut-être présent plusieurs fois lors d'une classification non supervisée de mots-clés d'une campagne. Afin de contrer ce problème, il convient de prendre en compte un changement d'échelle temporelle. Pour cela, nous proposons une agrégation de mots-clés d'un mois. Ainsi, on aura diminué au plus de 31 fois le nombre d'apparitions d'un même identifiant dans un mois.

Cette solution, que nous estimons préférable, s'avère utile. En plus, l'ajout de l'information au niveau des variables statistiques sera enrichi, et principalement la capacité de traitement sera réduite. En effet, suite à des analyses effectuées à partir des données (Quinn, 2011, Annexe 3), on remarque qu'en moyenne, pour la moitié des mots-clés dans chaque banque de données étudiée, on a un nombre de clics nul (Tableau 3.4).

De ce fait, nous considérons que l'échelle temporelle mensuelle génère de l'information suffisamment consistante pour être utilisée aux méthodes de classification non supervisée.

---

3 Présentation du logiciel R. Consulté le 20 avril 2014. Tiré de http://www.biostatisticien.eu/springeR/livreR_presentation.pdf

Tableau 3.4 : Répartition des clics et coûts dans une campagne publicitaire

Agence X		
groupe de mots-clés (centile)	Somme de clics par jour	Proportion du total (%)
95-100	78 667,12	80,32
90-95	7 542,28	7,70
85-90	4 167,52	4,26
80-85	2 728,48	2,79
75-80	1 901,27	1,94
70-75	1 329,85	1,36
65-70	896,39	0,92
60-65	542,36	0,55
55-60	161,45	0,16
50-55	0,00	0,00
45-50	0,00	0,00
40-45	0,00	0,00
35-40	0,00	0,00
30-35	0,00	0,00
25-30	0,00	0,00
20-25	0,00	0,00
15-20	0,00	0,00
10-15	0,00	0,00
5-10	0,00	0,00
0-5	0,00	0,00

Nous utilisons des fonctions d'agrégations telles que la somme et la moyenne pondérée pour effectuer ces changements.

Pour arriver à regrouper les mots-clés dans *n* classes pour chaque campagne, où *n* est fixe, une uniformisation des données est nécessaire. Il faut convertir les données à une échelle commune. C'est ce que nous aborderons dans la section suivante.

3.3.3.2 Normalisations des données

La différence d'amplitude des variables d'un mot-clé à un autre et d'une campagne à une autre peut poser des problèmes au niveau de la comparaison des groupes. Elle peut aussi avoir une forte influence sur les résultats de classification non supervisée. Cependant, une normalisation de données est nécessaire afin d'arriver à améliorer la qualité de traitement des données.

La fonction de normalisation devrait dépendre de la forme et de la concentration du nuage de points de données. Pour cela, on effectuera différentes approches de normalisations.

Type de normalisation

Les types de normalisation que nous avons tentés d'utiliser au cours de notre étude sont donnés par les formules suivantes:

Soit x_i une valeur de l'ensemble des données d'une variable quelconque *X*. On note la valeur normalisée de x_i par x_i' ;

- Normalisation centrée réduite

$$x_i' = \frac{x_i - \ moyenne(X)}{écart\ type(X)}$$

- Normalisations par la médiane

$$x_i' = \frac{x_i - \ médiane(X)}{\frac{1}{n}\sum_{i=1}^{n}|x_i - \ médiane(X)|}$$

La médiane est une alternative robuste pour l'estimation en présence de valeurs extrêmes. Elle peut être utilisée comme un estimateur du centre de la distribution des points.

- Normalisation par minmax

$$x_i' = \frac{x_i - \ \min(X)}{\max(X) - \min(X)} \qquad (1)$$

Celle qui a été retenue est la normalisation minmax. Ainsi, toutes les valeurs des variables seront réduites entre 0 et 1. Dans ce cas, une analyse de comparaison sera significative, sauf pour le cas de la variable position, déjà standardisée, où toutes les valeurs des positions dans diverses campagnes sont comprises entre 1 et 12.

Après les phases de traitement et de transformation de données qui pourraient nous amener à préciser notre modèle descriptif de données de campagnes publicitaires, nous exposons dans la section suivante la phase d'analyse de classification non supervisée par les méthodes de data mining. L'objectif de cette phase est de présenter les différents algorithmes utilisés pour déceler un nombre commun de classes de mots-clés dans chaque campagne publicitaire, soit à partir des patterns (visualisation) ou soit par des indices de performance de séparation des classes.

3.3.4 Classification non supervisée

Par rapport à ce que nous avons expliqué au chapitre précédent, la classification non supervisée est une phase de regroupement d'individus en différents groupes. La représentation des individus a pour but de découvrir un nombre de classes souhaité selon les comportements des mots-clés qui correspondent le plus souvent à des critères de proximité. La proximité des individus est mesurée par une fonction de distance entre chaque paire d'individus. Nous avons suggéré comme fonction la distance euclidienne ayant la formule suivante :

$$\mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

où x_i et y_i sont respectivement les coordonnées des vecteurs $\boldsymbol{x}$ et $\boldsymbol{y}$.

Il y a eu, certes, d'autres distances utilisées à travers des expériences de classification sur des données, nous permettant ainsi d'avoir une idée pour choisir la mesure la plus pratique. Comme exemple, citons la distance Manhattan définie par

$$\mathrm{d}(\boldsymbol{x}, \boldsymbol{y}) = \sum_i |x_i - y_i|.$$

Quant à l'information critique (nombre de classes), elle sera utilisée a priori par les algorithmes de classifications. À cette étape de classification non supervisée, le nombre de classes est inconnu et la plupart des algorithmes exigent ce paramètre. Pour cela, et dans le but de comparer les résultats, nous avons examiné plusieurs scénarios en générant un nombre de classes variant de 2 à 10.

Les méthodes de clustering dépendent généralement du choix initial des individus, ce qui peut exclure la certitude d'atteindre l'optimum global. Elles ne détectent pas bien les silhouettes de nuage de points en formes non sphériques (Baya & Granitto, 2013).

Pour un bon partitionnement, il convient à la fois de :

- minimiser la dispersion intra-classe afin d'obtenir une homogénéité dans chaque groupe.
- maximiser la dispersion inter-classe afin d'obtenir une hétérogénéité entre les groupes.

Pour ce faire, on utilisera les algorithmes suivants :

3.3.4.1 k-moyennes (k-means)

Développé par MacQueen (1967), k-means est l'un des algorithmes les plus connus en classification automatique. Toutefois, il exige un paramètre critique : le nombre de classes à générer. On part d'une partition arbitraire en *K* classes que l'on améliore itérativement, par une fonction de minimisation de la somme des distances entre chaque objet et le centre de son cluster, jusqu'à convergence du critère choisi. Et comme la partition finale pourrait dépendre de l'initialisation, on procède alors au traitement de façon répétitive. Chaque cluster est représenté par son centre appelé centroïde.

Notons que le problème de k-means est du type NP-difficile (Aloise, Deshpande, Hansen & Popat, 2009). Dans ce cas, il va de soi que l'on fasse appel aux méthodes heuristiques.

Ci-dessous, les détails de l'algorithme :

Algorithme 1 : k-moyennes (k-means)

Input : $E = (e_1, e_2, e_3, \dots, e_n)$ jeu de données de dimension *n* à regrouper

k (nombre de classes)

MaxIter (nombre d'itérations maximum)

Output : $C = \{c_1, c_2, \dots, c_k\}$ (ensemble des centres de classes)

$L = \{l(e_i) \,/\, i = 1,2, \dots, n\}$ (ensemble de données *E* libellées)

Initialisation :

Pour tout $c_k \in C, \quad c_k \leftarrow e_j \in E$ (ex : choisir aléatoirement *k* centres dans *E*)

iter = 0

Répéter jusqu'à (*iter* > *MaxIter*)

- Pour tout $e_i \in E$, a*ffecter* e_i *au centre le plus proche de* C
- *recalculer les centres de classes* c_k
- *arrêter si aucun élément ne change de groupe*
- $iter \leftarrow iter + 1$

Fin

3.3.4.2 PAM

Développé par Kaufman et Rousseeuw (1990), PAM n'est qu'une variante de l'algorithme de k-means. Il a l'avantage d'être plus robuste aux points atypiques ou valeurs aberrantes et il recourt aux médianes plutôt qu'aux moyennes. En effet, les centres de classes, appelés des médoides, se calcule par la médiane et non par la moyenne. Chaque cluster est représenté par son médoide qui en est l'objet.

Les détails de l'algorithme :

Algorithme 1 : k-médoides (PAM)

Input : $E = (e_1, e_2, e_3, \dots, e_n)$ jeu de données de dimension *n* à regrouper

k (nombre de classes)

Output : $M = \{m_1, m_2, \dots, m_k\}$ (ensemble des médoides)

$L = \{l(e_i) \,/\, i = 1,2, \dots, n\}$ (ensemble de données *E* libellées)

Initialisation :

Pour tout $m_k \in M, \quad m_k \leftarrow e_i \in E$ (choisir aléatoirement *k* médoides dans *E*)

Répéter

- Pour tout $e_i \in E \backslash M$, a*ffecter e_i au médoide le plus proche de M*
- *sélectionner un point non médoide e_i et un médoide m_k*
- *calculer le coût de remplacement de m_k par e_i*

(ex : critère d'erreur global $EG = \sum_{i=1}^{n} \sum_{e_i \in c_k} d(e_i, m_k)^2$; où c_k classe du médoide *k*)

- *permuter le médoide m_k par e_i si le coût est négatif*
- *arrêter si aucune amélioration n'est possible*

Fin

3.3.4.3 CLARA « Clustering LARge Application »

Également introduit par Kaufman et Rousseeuw (1990), l'algorithme CLARA est une variante de l'algorithme PAM où il a l'aptitude de traiter les grandes bases de données contrairement à PAM. Vu que nous opérons sur plusieurs échantillons en même temps, l'application de l'algorithme CLARA peut s'avérer avantageuse.

3.3.4.4 Fuzzy C-means

Développé par Bezdek (1981), Fuzzy C-means est un algorithme de classification floue non supervisée. Il introduit la notion floue dans le cadre où il autorise le chevauchement des classes. Ce chevauchement explique qu'un objet peut appartenir à une classe avec un certain degré d'appartenance compris entre 0 et 1. L'objet sera associé à la classe dont le degré d'appartenance est le plus élevé et les classes obtenues ne sont pas forcément disjointes.

Le principe est le suivant[4] :

[4] Tiré de http://www.sersc.org/journals/IJEIC/vol4_Is1/1.pdf

Minimiser la fonction du critère définie par :

$$J_s = \sum_k \sum_j (u_{jk})^s d^2(e_j, c_k)$$

sous les contraintes suivantes :

$$\sum_j u_{jk} = 1 , \qquad \forall\, k$$

$$u_{jk} \in [0,1] \quad \forall\, j, k$$

où

- u_{jk} le degré d'appartenance de l'objet *j* à la classe k, c_k le centroïde de la classe k et e_j le j ième objet.
- $d(e_j, c_k)$ est la distance entre e_j et c_k (généralement définie par la distance euclidienne).
- *s* le paramètre de réglage du degré de floue des classes.

Les étapes de l'algorithme sont les suivantes:

Algorithme 1 : Fuzzy c-means

Input : $E = (e_1, e_2, e_3, \dots, e_n)$ jeu de données de dimension *n* à regrouper

k (nombre de classes)

Output : $C = \{c_1, c_2, \dots, c_k\}$ (ensemble des centres de classes)

$L = \{l(e_i) \,/\, i = 1,2, \dots, n\}$ (ensemble de données *E* libellées)

Initialisation :

Choix arbitraire d'une matrice d'appartenance u_{kj}

Répéter

- *calculer les centroïdes des classes* c_k

$$c_k = \frac{\sum_j (u_{jk})^s . e_j}{\sum_j (u_{jk})^s}$$

- *réajuster la matrice d'appartenance suivant la position des centroïdes*

$$u_{jk} = \frac{\left(\frac{1}{d^2(e_j, c_k)}\right)^{1/(s-1)}}{\sum_j \left(\frac{1}{d^2(e_j, c_k)}\right)^{1/(s-1)}}$$

- *Calculer le critère de minimisation* J_s
- *arrêter s'il y a convergence*

Fin

Pour toutes les méthodes, on fait varier les paramètres pour avoir plusieurs partitions avec des nombres différents de classes. La qualité d'une méthode de clustering est évaluée par son habilité à découvrir les patterns cachés.

3.3.4.5 CLUES « CLUstering based on local Shrinking»

Introduit par Wang et al. (2007), l'algorithme CLUES est une approche basée sur le rétrécissement local des centroïdes pris aléatoirement au départ. L'algorithme Clues est une méthode heuristique qui cherche une solution locale. L'algorithme est composé en trois procédures principales : le rétrécissement, le partitionnement et la détermination du nombre optimal K voisin. La procédure de rétrécissement est influencée par l'approche du K voisin les plus proches. La valeur K commence par un petit nombre de points et augmente progressivement jusqu'à ce que la mesure d'indice de Silhouette soit optimisée. Nous n'allons pas décrire les détails de l'algorithme, mais il est conçu pour traiter des grandes bases de données en offrant un indice de qualité de classification. Il existe dans le package dans R.

Finalement, il est difficile d'analyser des volumes importants de données afin d'en tirer des informations utiles. D'où l'utilisation d'indicateurs de performance pour la classification.

3.3.4.6 Indice de qualité de classification

Comme indice de performance de classification, on a suggéré l'indice Silhouette, proposé par Kaufman et Rousseeuw (1990). Il est défini comme suit :

$$Silhouette = \frac{1}{n}\sum_{i=1}^{n} S(x_i)\,,$$

$$o\grave{u} \quad S(x_i) = \frac{b(x_i) - a(x_i)}{\max(a(x_i), b(x_i))}\,,$$

où $a(x_i)$ est la distance moyenne d'un point de donnée x_i par rapport aux autres points dans une classe donnée C et $b(x_i) = \min_{A \neq C} d(x_i, A)$, avec $d(x_i, A)$ = distance moyenne d'un point de données x_i dans classe C par rapport à tous les points de données dans la classe A.

L'interprétation de l'indice $S(x_i)$ est selon les valeurs suivantes :

Proche de 1 : le point x_i est bien classé

Proche de 0 : le point x_i est sur une limite entre les clusters

Proche de -1 : le point x_i est dans la mauvaise classe.

Estimation du nombre de classes

Après les tests de classification non supervisée sur l'ensemble d'échantillonnage, il reste le problème de la validation de nombre de clusters obtenus. Pour cela, on observe les résultats par rapport à la fréquence du nombre de clusters et on s'intéresse aux indices de qualité de classification appropriés pour estimer le nombre de classes. Pour y arriver, nous avons combiné les deux indices (fréquence et indice de qualité) afin de donner un score au résultat du nombre de classes obtenu. Considérons l'exemple suivant :

Tableau 3.5: Classement des nombres de groupes

Nombre de classes	Fréquence (f_i)	Indice Silhouette (*Sil*)	*fi * Sil*	Score
2	25	0,35	8,75	1
3	12	0,38	4,56	3
4	10	0,46	4,60	2

Illustrons par un schéma (Figure 3.6) notre processus de validation des classes:

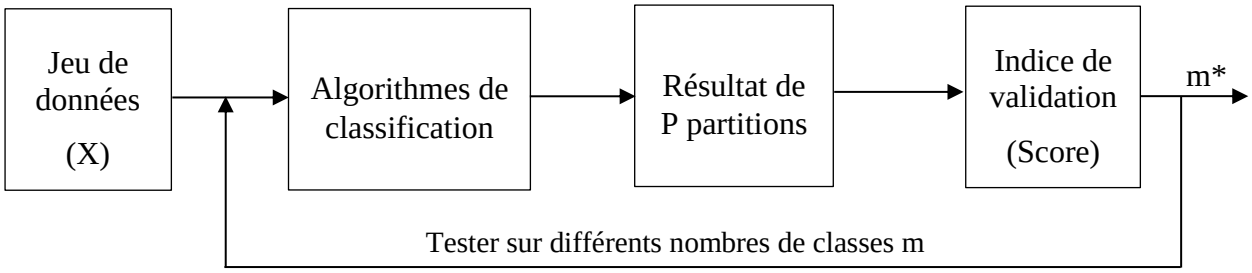


Figure 3.6: Processus de validation du nombre de classes

Il sera essentiellement question de déterminer les variables qui expliquent la séparation des mots-clés. Nous avons utilisé comme application pour cette analyse la fonction « Stepclass »[5] disponible dans le package « klaR » de R. Elle utilise la sélection des variables pas à pas par la méthode d'analyse discriminante linéaire afin de déterminer la contribution des variables qui expliquent l'appartenance des individus à des groupes. Il y a d'autres algorithmes proposés par Witten et Tibshirani (2010) ainsi que Friedman et Meulman (2004) (implémenté dans le package « sparcl » de R) pour la sélection des variables. L'application de ces algorithmes sur des ensembles de données montre que les mesures d'importance variables peuvent être améliorées de manière significative par itérations.

Une fois que le nombre de groupes est connu, un classement des mots-clés est requis, d'où la classification supervisée.

[5] http://cran.r-project.org/web/packages/klaR/klaR.pdf

3.3.5 Classification supervisée

Dans cette section, nous procéderons au classement des mots-clés d'une campagne publicitaire par différents tests de méthodes de classification supervisée.

Tel qu'expliqué à la section 2.3.2, à partir d'une base donnée d'apprentissage, dans la méthode de classification supervisée une estimation est faite sur le classement des mots-clés.

Le choix des méthodes de classification supervisée dépend de la stratégie d'agrégation. En effet, une stratégie d'affectation doit être en concordance avec la stratégie d'agrégation[6]. Pour cela, nous avons utilisé l'algorithme k-means qui a donné de meilleurs résultats et a prouvé sa performance quant au temps de traitement.

Toutefois, les résultats expérimentaux de la méthode de classification non supervisée ont décelé un nombre de classes égal à deux, c.-à-d. un classificateur de dimension 1. Ce qui n'est pas trop attendu et devient un cas simple linéairement séparable pour la classification supervisée. En effet, suivant une valeur de la variable séparatrice on classe tout simplement les données. Il nous reste donc à déterminer cette valeur par les méthodes de classification supervisée qui nous amène à affiner la répartition. On y reviendra plus en détail dans le chapitre suivant.

Par la suite, nous établirons les fonctions génériques des mots-clés correspondant aux deux classes découvertes.

3.3.6 Fonctions génériques

Comme nous l'avons mentionné ci-dessus, l'objectif principal de notre étude est d'améliorer les fonctions génériques établies par Quinn (2011), celles des prédictions des Clics et des CPC par rapport à la position. Et en nous appuyant sur l'étude existante d'analyse de données (Quinn, 2001), nous préservons les conditions déjà établies par Quinn pour l'extraction d'un ensemble de mots-clés permettant d'offrir des fonctions de prédiction de qualité.

---

[6] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_KMeans_Deploiement.pdf

Les critères (Quinn, 2001, p. 87) définis sont :

- écart-type des valeurs de position doit être supérieur à 1,5
- max (valeur de position) – min (valeur de position) ≥ 4
- nombre de jours de données disponibles dans les 120 derniers jours ≥ 100
- nombre moyen de clics par jours ≥ 20
- valeur de position minimal ≤ 5
- coefficient de détermination de la régression linéaire (R^2) ≥ 0,30
- la pente de la fonction régression des Clics par rapport à la position est négative
- la pente de la fonction régression des CPC par rapport à la position est négative

Ainsi, sur 20 agences, nous avons obtenus 436 mots-clés pour la prédiction des clics et 391 mots-clés pour la prédiction des CPC. Ce sont des nombres apparemment très limités, néanmoins nous nous contentons des critères mentionnés ci-dessus, jugés nécessaires afin de fournir des fonctions de régressions acceptables.

Les équations (Quinn, 2011, p. 101) des fonctions de prédictions des variables clic et CPC ont les formules suivantes :

$$clic = k_{clic} * (c_{clic})^{pos} \quad (2)$$

$$cpc = k_{cpc} * (c_{cpc})^{pos} \quad (3)$$

où :

- k_{clic} et k_{cpc} sont des paramètres constants propres à chaque mot-clé caractérisant l'échelle de grandeur de chaque courbe.

- c_{clic} et c_{cpc} sont des constantes globales pour tous les mots-clés, caractérisant le taux de décroissance des fonctions qui sont relativement les mêmes d'un mot-clé à l'autre (pentes exponentielles des variables Clic et CPC par rapport à la position).

- clic, cpc et pos sont les variables respectives de Clic, CPC et position.

Les paramètres globaux c_{clic} et c_{cpc} sont estimés par la moyenne des valeurs observées du taux de décroissance des clics et des CPC calculés sur l'ensemble des mots-clés obtenus par les critères exigés. Le calcul sera détaillé dans le chapitre suivant.

En considérant le résultat obtenu de la classification non supervisée, c.-à-d. le nombre de classes égal à deux, nous procédons au raffinement des fonctions de prédictions par l'amélioration des valeurs des paramètres constants, et cela, pour chaque groupe. En effet, chaque classe aura ses propres fonctions génériques. Dans ce cas, nous serons appelés à déterminer quatre paramètres constants dans chaque groupe, dont deux constantes globales et deux constantes propres au mot-clé qui sont $c_{clic}^{[1]}$ et $c_{cpc}^{[1]}$; $k_{clic}^{[1]}$ et $k_{cpc}^{[1]}$ pour le groupe 1 et $c_{clic}^{[2]}$ et $c_{cpc}^{[2]}$; $k_{clic}^{[2]}$ et $k_{cpc}^{[2]}$ pour le groupe 2.

Par ailleurs, certain critères cités ci-dessus seront proportionnellement ajustés d'après les limites des classes obtenues.

Pour arriver à une bonne estimation de ces paramètres, notamment celles des constantes propres au mot-clé, on doit déterminer ces derniers sur l'enrichissement de l'information. En effet, les paramètres k_{clic} et k_{cpc} , qui représentent respectivement les valeurs à l'origine des fonctions génériques Clic et CPC, sont déterminés en utilisant les points moyens des observations de coordonnées $(posMoy, clicMoy)$ et $(posMoy, cpcMoy)$ tels que :

$$posMoy = \frac{1}{n}\sum_{i=1}^{n} pos_i$$

$$clicMoy = \frac{1}{n}\sum_{i=1}^{n} clic_i$$

$$cpcMoy = \frac{1}{n}\sum_{i=1}^{n} cpc_i$$

qui sont calculés par la suite à partir des équations (2) et (3) :

$$k_{clic} = \frac{clicMoy}{(c_{clic})^{posMoy}} \quad (4)$$

$$k_{clic} = \frac{cpcMoy}{\left(c_{cpc}\right)^{posMoy}} \quad (5)$$

Notons que le calcul de ces paramètres pour chaque groupe sera le même, en tenant compte seulement de la séparation des observations du mot-clé par rapport aux classes.

Une fois que les formules sont établies, on procède à une série de tests sur les mots-clés de diverses campagnes tout en augmentant les périodes de 30, 60 à 90 jours. Il suffit de voir lorsqu'on augmente l'information, si notre modèle s'enrichit. Pour cela, des tests de validation de paramètres seront en étude dans la section suivante.

3.3.7 Validation des paramètres

L'inférence statistique classique, développée pour traiter des petits échantillons, ne fonctionne plus pour les très grands ensembles de données, où paradoxalement tout devient significatif (Saporta, 2004). Dans ce cas, des tests par validation sont suggérés (Clément, 2013). Pour effectuer ces tests, nous avons choisi un ensemble de jeux de données sur une période de 180 jours, constitués de *n* enregistrements de mots-clés. Il serait judicieux d'opter pour des données récentes. Le partitionnement de l'échantillon d'un jeu de données, selon les recommandations des auteurs, se constitue de manière à ce que les deux tiers des données soient affectés à l'apprentissage sur lequel on estime les paramètres en question. Le tiers restant est utilisé pour valider le comportement du modèle (voir Figure 3.7). On compare la qualité de prédiction sur une période de 30 jours consécutifs.

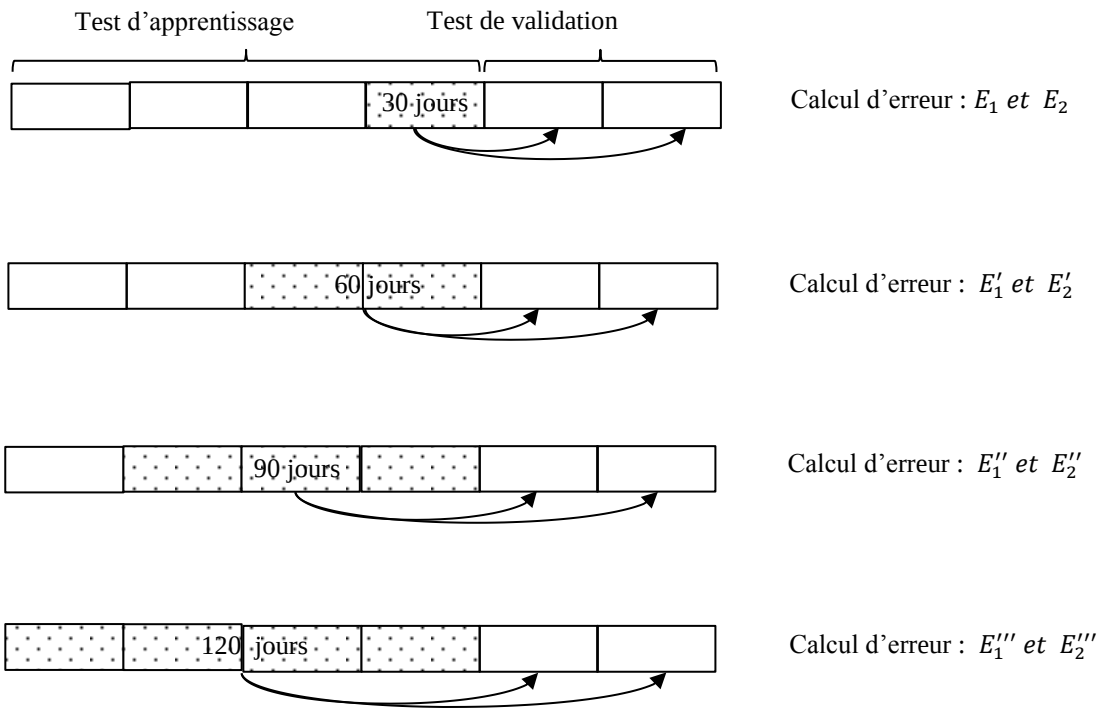


Figure 3.7: Partition de l'échantillon pour test

On remarque qu'il n'y a pas eu vraiment de tests de validation croisée du moment qu'on doit respecter la continuité du temps (le passée prédit le futur).

Tel qu'illustré sur la Figure 3.7, on estime les paramètres sur les données d'apprentissage (ceux en pointillé), puis on teste le modèle sur les données de validation qui sont constituées en deux périodes différentes de 30 jours. Les écarts observés (ex : E_1 et E_2) de chaque jeu données seront quantifiés par différentes mesures.

Indicateurs de mesure d'écarts

L'indicateur de mesure d'écarts nous informe sur l'estimation d'une erreur globale de notre échantillon choisi. Afin de déterminer cet estimateur, on calcule d'abord les écarts rapportés entre les valeurs observées et les valeurs prédites pour chaque mot-clé, associés à chacune des fonctions de prédictions Clic et CPC par différentes mesures qui quantifient les erreurs.

La mesure suggérée est celle de la moyenne d'erreur absolue (connue sous l'acronyme MAE) qui est un indicateur pratique de comparaison. Elle est moins affectée par les erreurs de prédiction les plus importantes (en particulier la variable Clic où sa valeur augmente de façon exponentielle en position prémium).

$$MAE = \frac{1}{N}\sum_{j=1}^{N}\left|y_j - \hat{y}_j\right| \qquad (6)$$

où y_j est la valeur observée au jour j, $\hat{y}_j$ la valeur prédite au jour j et N le nombre de jours d'observations.

Ainsi dans le cas de notre partitionnement, le calcul d'erreur total sera la somme de la moyenne d'erreurs absolue de prédictions de chaque groupe :

$$SMAE = MAE^{[1]} + MAE^{[2]}$$

où $MAE^{[1]}\ et\ MAE^{[2]}$ désignent respectivement les moyennes d'erreurs absolues de prédictions du groupe 1 et 2.

Finalement, les deux estimateurs de l'erreur globale de l'échantillon, associés aux fonctions de prédictions Clic et CPC, seront la somme des mesures de la moyenne d'erreur absolue de chaque mot-clé pondérée par son nombre d'observations sur une période de 30 jours.

1- Estimateur d'erreur globale associé à la prédiction des clics

$$EG_{clic} = \frac{\sum_k n_k * SMAE\_clic_k}{\sum_k n_k} \qquad (7)$$

2- Estimateur d'erreur globale associé à la prédiction des CPC

$$EG_{cpc} = \frac{\sum_k n_k * SMAE\_cpc_k}{\sum_k n_k} \qquad (8)$$

$SMAE\_clic_k$ est la somme de moyennes d'erreurs absolues du mot-clé k associée à la prédiction des clics;

$SMAE\_cpc_k$ est la somme de moyennes d'erreurs absolues du mot-clé k associée à la prédiction des CPC;

et n_k le nombre d'observations du mot-clé k sur une période de 30 jours.

Suite aux résultats des estimateurs d'erreurs globaux de l'ensemble de jeux de données (on procède à chaque jeu de données selon la Figure 3.7); nous évaluons la meilleure période d'apprentissage afin de calculer les bonnes valeurs des paramètres k_{clic} et k_{cpc} .

Une fois que les paramètres de chaque groupe sont établis, une comparaison des fonctions génériques sera décrite à la section suivante.

3.3.8 Comparaison des fonctions génériques

Après avoir déterminé les paramètres des nouvelles fonctions génériques des mots-clés associés à chaque classe, nous procédons à la comparaison de ces fonctions génériques associées aux classes avec celles établies par Quinn. Cela a pour but d'évaluer la performance de notre méthode.

La méthode est la même que celle vue à la section précédente. Il suffit de calculer la moyenne d'erreur absolue de prédictions (MAE), sur les données de chaque groupes, associée à la fonction (Clic; CPC) pour chaque mot-clé selon la fonction générique précédente. Ensuite refaire le même calcul avec les fonctions génériques actuelles. Il nous reste à évaluer la comparaison tout en calculant :

$$RCM = \frac{SMAE_{actuel}}{SMAE_{préc}} \qquad (9)$$

où $SMAE_{actuel} = MAE_{actuel}{}^{[1]} + MAE_{actuel}{}^{[2]}$

$SMAE_{préc} = MAE_{préc}{}^{[1]} + MAE_{préc}{}^{[2]}$

RCM est le rapport de comparaison de deux mesures de chaque mot-clé

$MAE_{préc}{}^{[j]}$ est la moyenne d'erreur absolue de la fonction de prédiction précédente calculée sur les données du groupe *j*, *j*=1,2.

$MAE_{actuel}^{[j]}$ est la moyenne d'erreur absolue de la fonction de prédiction actuelle du groupe j, j=1,2.

Ainsi pour un échantillon choisi, une mesure d'erreur globale pondérée est avantageuse afin de mieux comparer les résultats.

Les estimateurs de comparaison d'erreurs globales associés à la prédiction des clics et de CPC sont respectivement :

$$ECG_{clic} = \frac{\sum_k n_k * RCM\_clic_k}{\sum_k n_k} \quad (10)$$

$$ECG_{cpc} = \frac{\sum_k n_k * RCM\_cpc_k}{\sum_k n_k} \quad (11)$$

où

RCM_clic_k est le rapport de comparaison d'erreurs des fonctions génériques actuelles et précédentes du mot-clé k associée à la prédiction des clics;

RCM_cpc_k est le rapport de comparaison d'erreurs des fonctions génériques actuelles et précédente du mot-clé k associée à la prédiction des CPC;

et n_k le nombre d'observations du mot-clé k sur une période de 30 jours.

Si le rapport global est inférieur à 1, il y a eu une amélioration des fonctions génériques par la méthode de catégorisation sinon (le rapport est supérieur à 1) il n'y a pas eu d'amélioration selon cette catégorisation.

CHAPITRE 4 : EXPÉRIMENTATION ET RÉSULTATS

Dans ce chapitre nous allons décrire les étapes d'expérimentation selon la démarche méthodologique présentée précédemment, dont le contenu principal est le suivant:

- Présentation et préparation des données
- Extraction et prétraitement des données
- Classification non supervisée
- Classification supervisée
- Fonctions génériques et validation des paramètres
- Comparaison des fonctions génériques
- Dans chacune de ces étapes, les résultats obtenus sont présentés sous formes de tableaux ou de figures

4.1 Présentation et préparation des données

L'échantillon à étudier

Tel qu'expliqué à la section 3.3.2 (Échantillonnage, p. 30), l'échantillon ciblé est composé de campagnes de divers agences récoltant un nombre important de clics. En effet, si on extrait au hasard un échantillon de campagnes sans condition, on peut se retrouver avec des campagnes dont le nombre total de clics est inférieur à 10 pour un nombre d'observations qui dépasse les 500. Dans ce cas, les données de la variable clic seront presque toutes nulles. Le Tableau 4.1 qui résume les statistiques de campagnes publicitaires montre cette sélection. Ce tableau permet aussi de constater que la durée de campagne et le nombre d'observations jouent un rôle important sur le volume de clics.

Pour cela, la condition d'avoir un nombre important de clics aux campagnes est jugée nécessaire. Cette observation est également relevée par (Quinn, 2012, p. 83).

Rappelons également que cette base de données contient aussi des tests de campagnes expérimentales.

Tableau 4.1: Statistiques d'un échantillon de campagnes publicitaires

				Données totales (clic = 0 et clic > 0)									Données seulement avec clic > 0			
N°	Agency_ID	CampagnID	Durée de campagne	Nombre de mots-clés	Nombre d'observations	Impression Total	Position Pondérée	Clic Total	Coût Total ($)	Conversion total	Taux de clic par observation(%)	Taux de conversion (%)	Nombre de mots-clés	Nombre observation	Impression total	Position Pondérée
1	AA	0001	100	40	1 036	1 666	4	23	173,35	1	1,83	0,10	10	19	36	1
2	BB	0002	87	361	12 661	30 078	3	1 657	2 033,49	13	11,07	0,10	185	1 402	5 859	2
3	BB	0003	32	64	1 278	31 748	3	354	329,45	1	15,49	0,08	34	198	18 389	3
4	BB	0004	19	65	1 291	17 241	5	66	82,08	0	4,73	0,00	21	61	1 813	5
5	CC	0005	65	191	5 220	29 218	5	40	10,91	0	0,77	0,00	17	40	1 015	5
6	DD	0006	38	188	12 373	1 489 318	4	15 059	19 405,67	0	23,46	0,00	109	2 903	1 376 094	4
7	EE	0007	693	748	121 711	1 469 810	3	52 882	23 161,56	380	20,97	0,31	522	25 524	996 426	2
8	FF	0008	145	51	6 210	335 870	5	1 055	913,08	0	12,25	0,00	48	761	120 120	4
9	FF	0009	26	28	269	1 155	15	1	0,94	0	0,37	0,00	1	1	6	13
10	FF	0010	311	96	21 073	2 351 442	5	15 841	7 419,96	154	30,98	0,73	92	6 529	1 983 261	5
11	GG	0011	1	98	134	673	8	4	2,75	0	2,24	0,00	3	3	46	9
12	GG	0012	1	103	155	1 119	7	0	0,00	0	0,00	0,00	0	0	0	0
13	GG	0013	10	262	794	4 837	9	5	2,32	0	0,63	0,00	4	5	292	8
14	GG	0014	10	184	612	904	10	4	1,32	0	0,49	0,00	3	3	5	5
15	GG	0015	10	238	1 401	14 250	8	23	16,89	0	1,43	0,00	15	20	1 069	7
16	GG	0016	15	533	5 193	93 353	7	96	96,78	0	1,56	0,00	45	81	7 596	6
17	GG	0017	20	537	3 773	146 160	4	50	33,11	0	0,95	0,00	21	36	82 963	4
18	GG	0018	21	364	2 530	6 354	5	51	21,63	6	1,70	0,24	22	43	345	4
19	GG	0019	21	670	3 771	19 290	4	81	28,47	6	1,43	0,16	29	54	10 426	4
20	GG	0020	90	222	7 477	51 280	8	212	125,17	10	2,35	0,13	52	176	5 896	7
21	GG	0021	71	31	408	2 702	7	144	65,36	3	7,60	0,74	5	31	1 253	3
22	HH	0022	1128	107	38 991	390 705	4	5 498	5 220,61	14	11,53	0,04	87	4 495	127 492	4
23	HH	0023	1128	157	39 155	763 086	4	5 182	5 119,45	7	10,45	0,02	109	4 092	250 535	4
24	HH	0024	1128	180	61 782	480 581	4	7 805	5 469,88	16	9,93	0,03	146	6 133	167 491	3
25	JJ	0025	36	66	1 189	12 372	4	13	56,43	0	1,01	0,00	5	12	577	4
26	JJ	0026	36	50	588	3 384	2	7	33,18	0	1,19	0,00	5	7	305	2
27	JJ	0027	36	65	786	6 248	3	10	47,48	0	1,27	0,00	7	10	324	2
28	JJ	0028	228	303	35 914	246 738	3	1704	2 269,18	0	3,80	0,00	108	1 365	31 320	3
29	JJ	0029	38	91	2 222	43 932	2	51	214,56	0	2,21	0,00	17	49	3 356	2
30	JJ	0030	288	165	16 168	186 563	3	424	2 627,18	0	2,41	0,00	54	389	28 475	3
31	JJ	0031	289	178	8 973	61 380	4	74	350,79	0	0,80	0,00	32	72	1 285	4
32	JJ	0032	227	358	104 922	672 644	3	8740	11 351,30	0	6,00	0,00	278	6 296	133 287	3
33	JJ	0033	45	417	6 950	29 322	3	510	792,97	0	5,67	0,00	80	394	4 554	2
34	JJ	0034	199	259	33 726	350 739	6	5 637	2 972,00	0	4,41	0,00	134	1 489	140 152	6
35	JJ	0035	147	267	5 674	33 989	4	723	513,17	0	10,22	0,00	56	580	8 879	3

En outre, on a testé des campagnes à faible nombre de clics lors de l'application de la classification non supervisée. Ces tests n'ont pas donné de résultats significatifs. Vu le nombre de tests effectués, on ne peut pas décrire toutes les procédures pratiquées dans ce mémoire. On décrit uniquement les étapes de la réalisation qui ont menés au résultat final tout en expliquant les obstacles apparus durant la démarche.

Le choix, de notre échantillon de campagnes, est fait aléatoirement par un algorithme sous forme de langage R qui extrait les informations à partir du serveur de base de données d'Acquisio en langage de requête SQL. On a choisi les campagnes dont le nombre d'observations est élevé. On a estimé à 200 la taille de notre échantillon de campagnes publicitaires. Cela est insuffisant, vu la taille de notre population, mais on a jugé ceci acceptable vu que le temps de traitement du nombre important de mots-clés qui leurs sont associés est trop fastidieux.

Devant ces contraintes, nous concluons que notre échantillon de 200 campagnes publicitaires, composé de plusieurs types de domaines est significatif. Ces campagnes sont réparties sur 30 agences dont le nombre d'observations de mots-clés dépasse 15 000. Cet échantillon est le plus représentatif pour toute campagne à performance (collectant un nombre important de clics) afin de s'assurer que les résultats de la classification non supervisée ne dépendent pas d'une campagne à une autre. À l'annexe A, nous présentons un extrait de modèle de notre échantillon synthétisé sous forme de tableaux.

Description de l'échantillon

Notre échantillon est décrit par les caractéristiques des mots-clés associés aux campagnes extraites de notre base de données.

Les variables observées et les variables mesurées, telles que mentionnées dans la section 3.3.1, qui seront utilisées pour la classification sont les suivantes :

- Identifiant
- Impression
- Position
- Clic
- Coût
- Conversion
- CPC
- CTR

– Compte mot

L'unité statistique est l'identification du mot-clé et la variable « *compte mot* » est une mesure comptabilisant le nombre de mots d'un mot-clé.

Notons que la variable conversion correspond à l'objectif réel visé par les campagnes publicitaires. Malheureusement, les valeurs observées de cette variable sont presque toutes nulles dans une campagne. Afin de ne pas corrompre nos résultats de classification non supervisée, la variable en question sera écartée (voir Annexe A, taux de conversion par observations). En Annexe A, on remarque que la variable conversion représente au maximum 3% d'informations au niveau d'une campagne.

Analyse des variables

Par une méthode de l'analyse en composantes principales (connu sous le sigle ACP[7]), qui résume en un aperçu nos variables quantitatives par un graphique global, ceci permettra de comprendre la structure des données analysées.

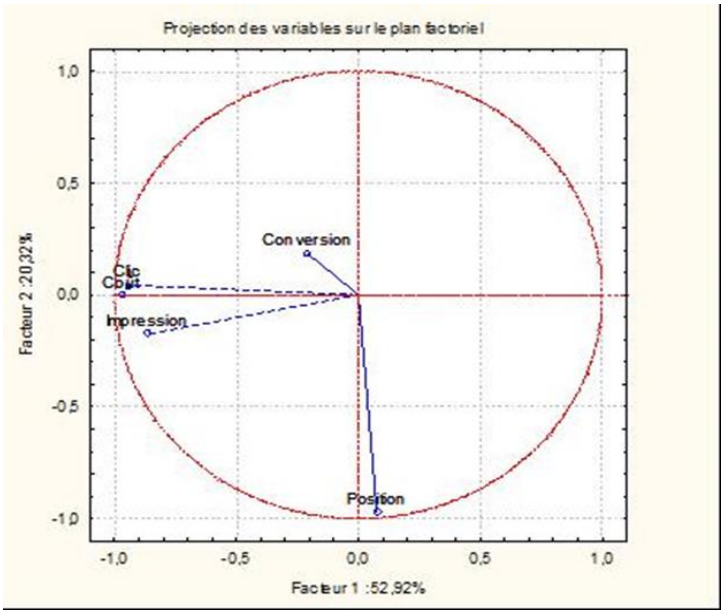


Figure 4.1: Analyse de variables par la méthode ACP

[7] Tiré de http://en.wikipedia.org/wiki/Principal_component_analysis

On remarque que la variable position semble opposée au reste des variables et que la représentation de la variable conversion pointe vers une autre direction et n'est pas assez éloignée du centre du cercle.

D'autre part, les variables impression, clic et coût semblent avoir certaines similitudes puisqu'elles sont orientées dans le même sens.

Relation entre les variables

Afin d'avoir un aperçu des relations entre les variables, la fonction « pairs » de R permet de visualiser, par un nuage de points, les relations deux à deux des variables qualitatives d'une campagne donnée (voir Figure 4.2). La visualisation est essentielle, elle permet une première approche des données, met en évidence la difficulté éventuelle du problème et oriente l'étude vers une telle technique (Bouveyron, 2013, p. 63).

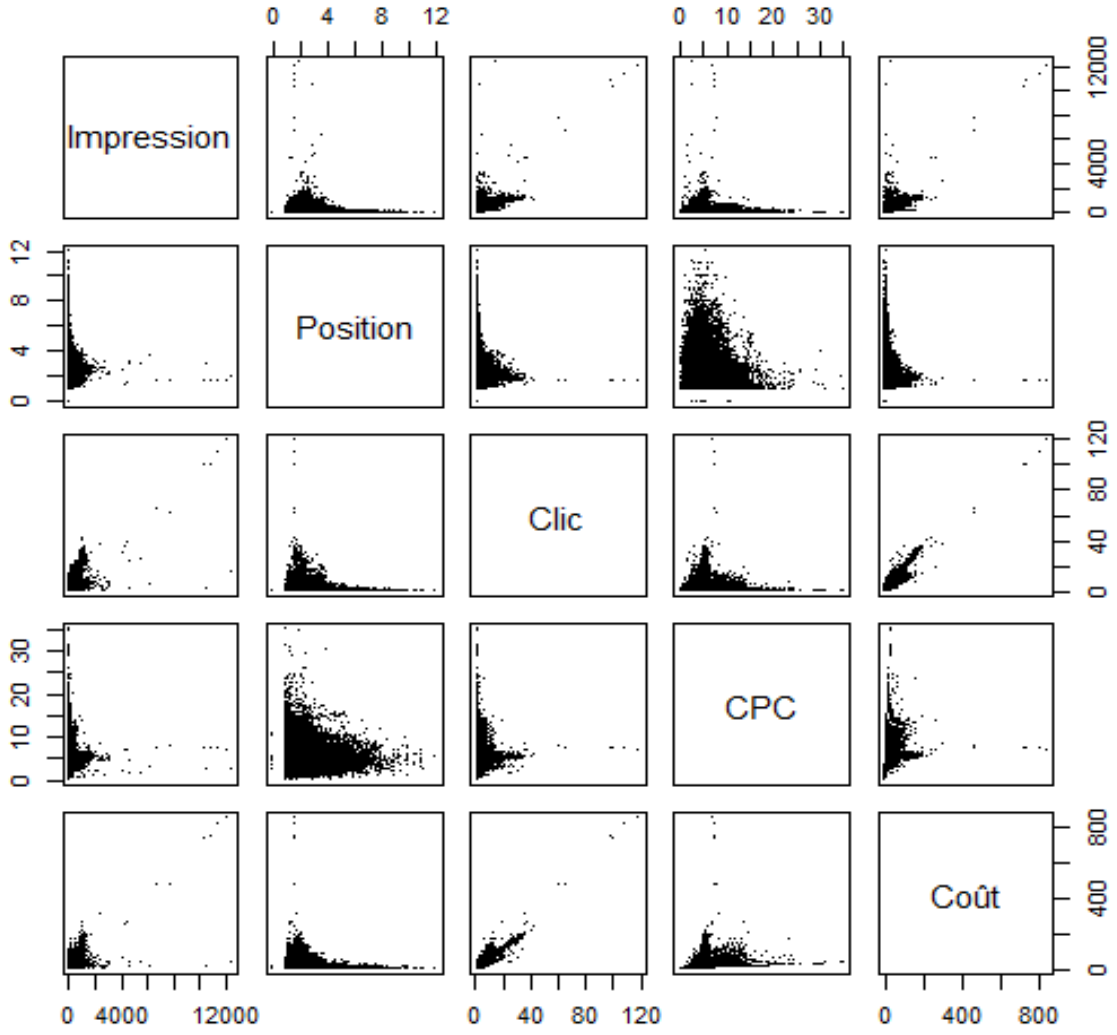


Figure 4.2: Nuages de points illustrant les relations entre les variables

Les points extremums existant comme valeurs aberrantes (Figure 4.2) ne seront pas écartés. En général, les points qui s'écartent grandement des autres points sont considérés des données aberrantes. Mais le fait qu'une annonce atteint une position top, on s'attend à avoir un maximum de clics. Il est donc important de préserver ses valeurs.

La prise en compte de ces valeurs, qui peuvent modifier la signification de la moyenne ou fausser l'analyse ultérieure de la classification, s'avère judicieuse.

On présente en Annexe B, quelques illustrations graphiques des relations de variables de divers campagnes afin d'avoir une idée sur le comportement des campagnes. On constate qu'il y a une certaine tendance de similitude de comportement entre les différentes campagnes ayant un nombre important d'observations. En effet, si on avait standardisé les variables de chaque campagne, la ressemblance serait meilleure. Nous avons réitéré ces analyses sur d'autres campagnes, on a observé approximativement les mêmes tendances de similitudes.

Ainsi l'hypothèse de l'uniformité de comportement des campagnes va nous aider à pouvoir généraliser le nombre de classes trouvées.

4.2 Extraction et prétraitement des données

À partir de ce qui a été mentionné dans la section 3.3.2 et 3.3.3, notre problématique est de gérer les données doublons pour une extraction efficace de l'information.

Notons dans un premier temps que la classification non supervisée de mots-clés de campagne de taille non réduite (i.e. observations de données sur une échelle quotidienne) n'a pas donné de résultats significatifs.

Nous agrégeons ces différentes informations des données quantitatives sur une échelle mensuelle par une somme et une moyenne pondérée pour le calcul de la variable position agrégée.

$$Pos.moy = \frac{\sum Impression * \ position}{\sum Impression}$$

Cette agrégation doit augmenter l'information au niveau des données et réduire la redondance des observations de données du mot-clé afin de pouvoir gérer les doublons dans la situation où il y aura le même mot-clé dans plusieurs classes.

Illustrons un exemple de données d'un mot-clé d'une campagne sur une durée d'un mois par le Tableau 4.2.

Cet exemple est extrait d'une campagne de durée de 12 mois qui a un total de 13 238 d'observations dont 11 121 valeurs nulles de clics; soit 16 % d'informations soutirées de cette matrice.

Tableau 4.2: Données agrégées d'un mot-clé

Données détaillées d'un mot-clé d'une campagne donnée sur une période d'un mois.

Jour	Clic	Coût	Impression	Position	Conversion
1	0	0	3	3	0
2	0	0	10	4,6	0
3	1	1,21	9	3	0
4	0	0	9	2,89	0
5	0	0	2	2	0
6	0	0	9	2,89	0
7	0	0	1	3	0
8	0	0	11	2,45	0
9	0	0	4	1,75	0
10	1	1,08	6	2,33	0
11	1	1,04	10	2,7	0
12	0	0	6	4,5	0
13	1	0,9	17	2,82	0
14	0	0	5	2,4	0
15	0	0	5	2,2	0
16	0	0	14	2,93	0
17	1	1,22	4	3	0
18	1	1,17	8	2,75	0
19	1	0,48	7	1	0
20	0	0	3	4	0
21	0	0	5	3,2	0
22	0	0	4	4,5	0
23	1	1,08	13	2,23	0
24	0	0	2	1,5	0
25	0	0	3	1,67	0
26	1	0,32	3	3	0
27	0	0	1	1	0
28	1	1,16	11	2,36	0
29	0	0	1	2	0
30	0	0	2	1,5	0

Mois	Somme Clic	Somme Coût	Somme Impression	Position pondérée	Somme Conversion
Données agrégées: 1	10	9,66	188	2,77	0

Le regroupement des données est fait au niveau de la base de données Oracle, ou le processus de traitement est plus rapide. Une fois les données regroupées, le résultat est redirigé vers le logiciel R.

Transformation des entrées

Comme expliqué dans la section 3.3.3 l'efficacité de généraliser les résultats de classification non supervisée de campagnes ne sera pas significative si on ne tient pas compte des variations de données entre les campagnes. Pour cela, un moyen de normalisation des données sera établi afin d'éliminer les différences qui existent entre les valeurs des mots-clés.

La normalisation minmax, définie par l'équation (1), est utilisée pour toutes les valeurs des variables qui seront réduites entre 0 et 1, sauf la variable position tel qu'expliqué dans la section 3.3.3.

Présentation du modèle

Après tous les traitements cités ci-dessus, le modèle d'une campagne donnée se présente sous la forme suivante :

Tableau 4.3: Modèle de campagne après traitements

	SommeImpression	PositionMoyPondérée	SommeClic	CPCmoy	SommeCoût	CTR	CompteMot
1	0,11663	2,95	0,13009	0,12892	0,07418	0,00550	2
2	0,22129	2,66	0,36364	0,14085	0,24754	0,00750	2
3	0,08375	2,21	0,07524	0,12736	0,04284	0,00450	2
4	0,08134	2,92	0,05799	0,11227	0,02784	0,00350	2
5	0,08467	3,57	0,06270	0,09066	0,02625	0,00350	2
6	0,12974	3,07	0,10972	0,11526	0,05778	0,00400	2
7	0,10079	2,91	0,07367	0,12515	0,04162	0,00350	2
8	0,08489	3,47	0,07053	0,09634	0,03172	0,00450	2
9	0,16231	3,09	0,10188	0,12457	0,05727	0,00300	2
10	0,03210	2,73	0,02194	0,09536	0,00991	0,00350	2
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
44545	0,03886	2,88	0,02194	0,13519	0,01357	0,00250	2
44546	0,11827	3,35	0,08464	0,08561	0,03341	0,00300	2
44547	0,04404	3,10	0,02351	0,07778	0,00814	0,00200	2
44548	0,02426	1,76	0,02194	0,14874	0,01426	0,00500	2
44549	0,00056	2,24	0,00157	0,15811	0,00111	0,01900	2
44550	0,00050	2,15	0,00157	0,12881	0,00090	0,02250	2
44551	0,00099	1,89	0,00157	0,19843	0,00139	0,00400	2

La figure ci-dessous illustre les relations de variables quantitatives de cette campagne de données.

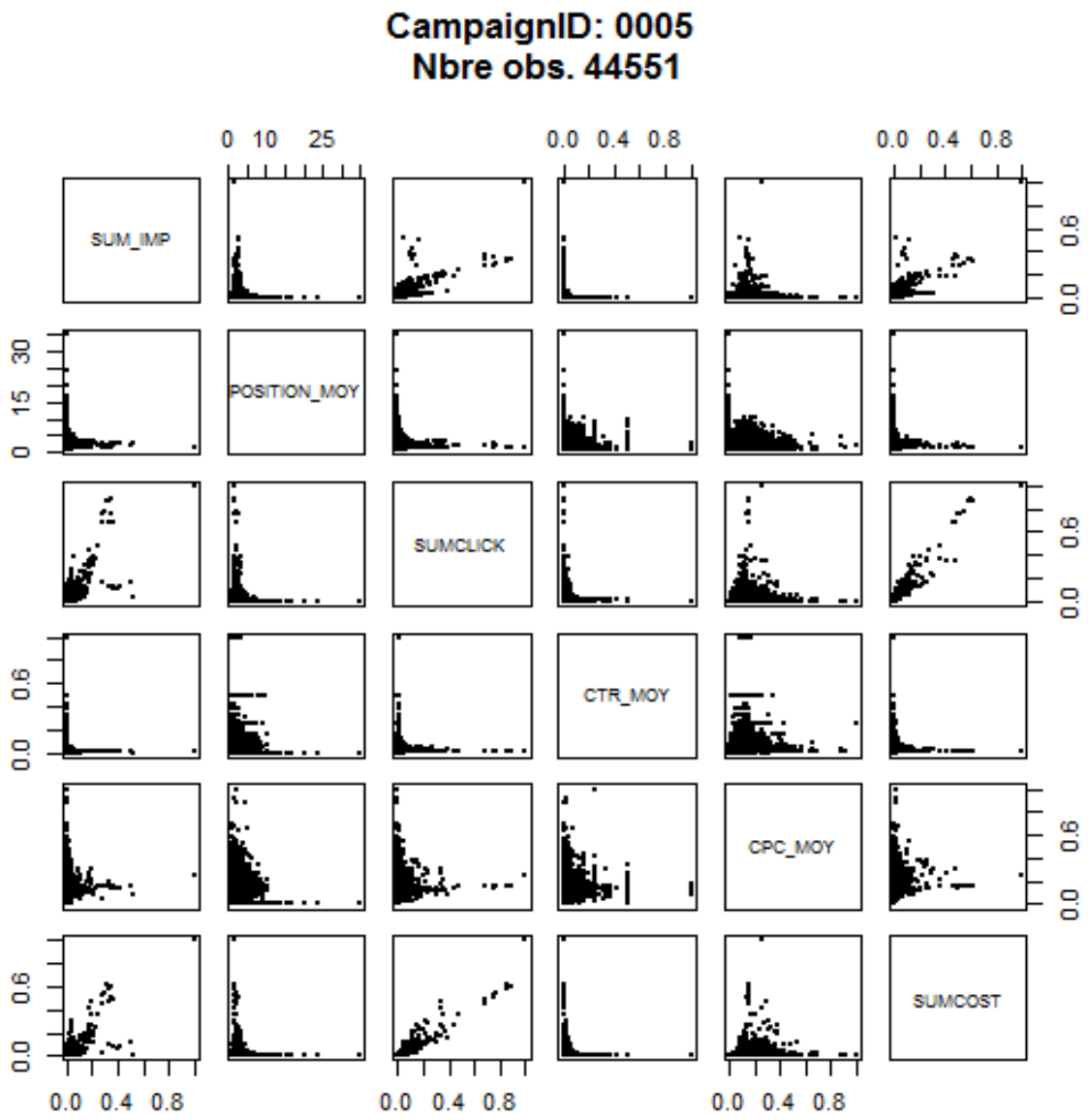


Figure 4.3: Nuage de points d'une campagne de données normalisées

La figure suivante nous présente les mêmes données de campagne avec la variable clic strictement positive.

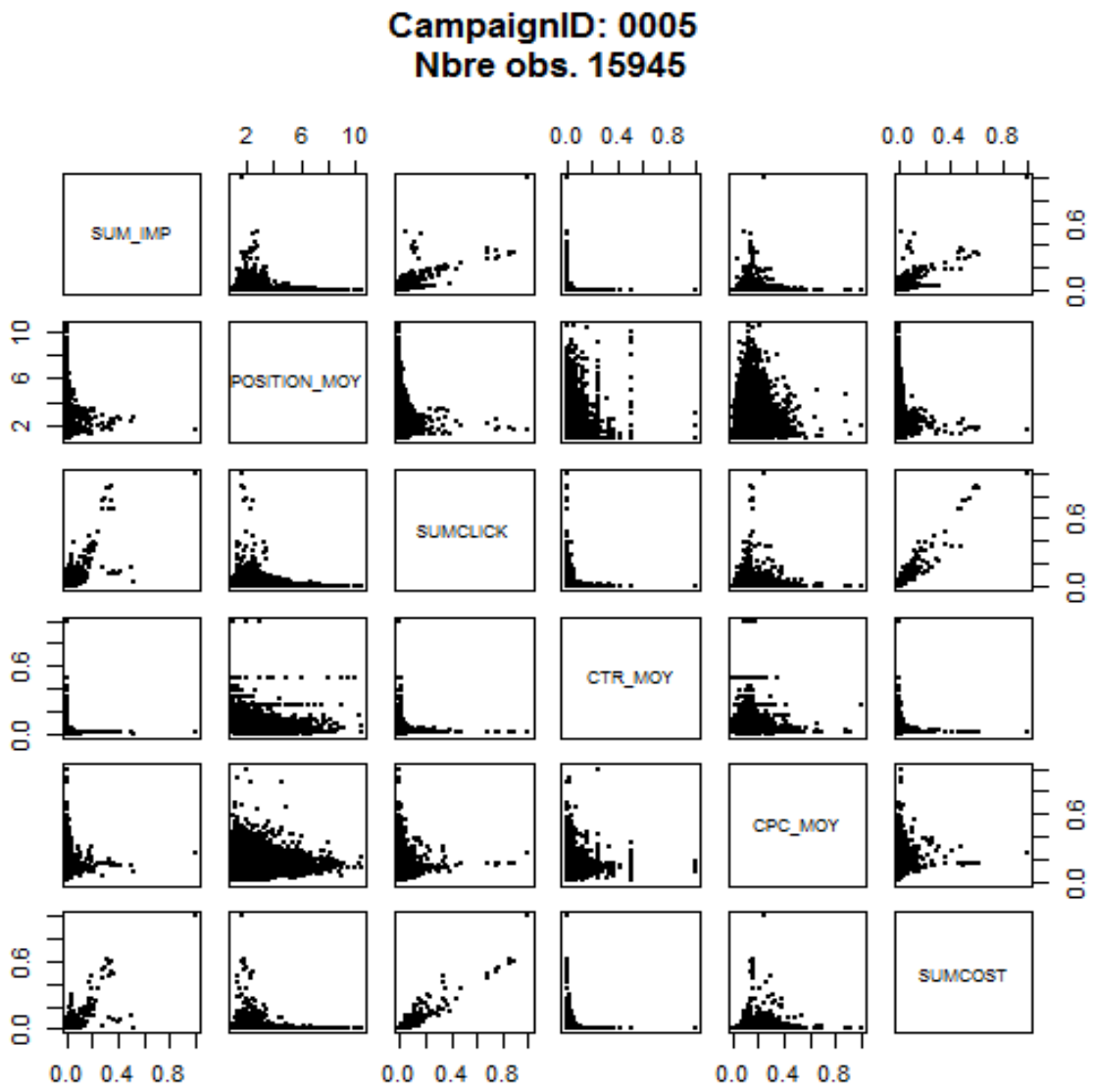


Figure 4.4: Nuage de points d'une campagne de données normalisées avec clic non nul

4.3 Classification non supervisée.

Étant donnée un échantillon d'apprentissage représentatif de 200 campagnes publicitaires dont leur nombre de classes n'étant pas connu, on espère arriver à déterminer par les méthodes de classification non supervisée un nombre fixe de classes pour toutes les campagnes. Pour cela, on

essaye de trouver une séparation standard qui se base sur la fréquence d'apparition du nombre de classes.

Par un schéma, nous tentons de représenter le problème pour une compréhension claire.

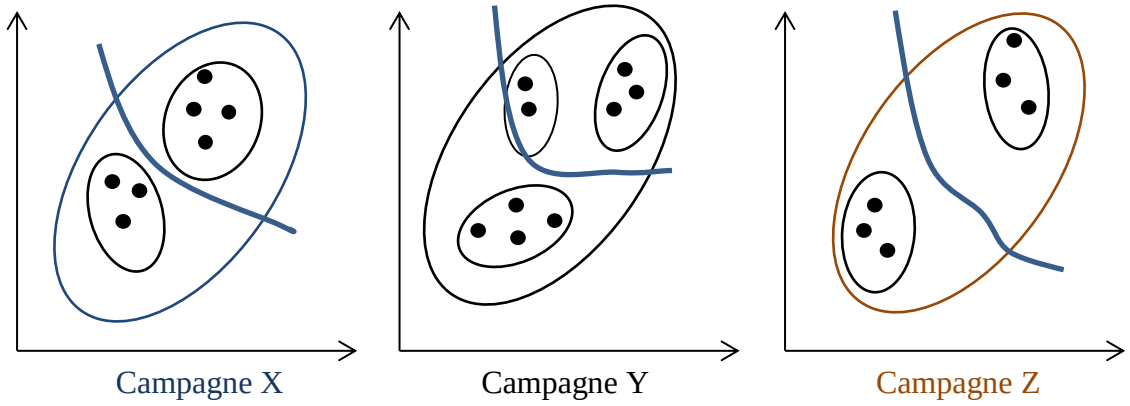


Figure 4.5: Exemple de classification pour une séparation standard

Dans cet exemple (Figure 4.5), où on montre que chaque campagne a sa propre classification, le nombre de classes standards égal à deux est le mieux proposé pour une généralisation sur l'ensemble de ces campagnes.

Le nombre de classes étant arbitraire, il intéressant dans une première étape de choisir un algorithme de classification automatique non supervisée pour avoir une idée sur le nombre de classes qui entourent les campagnes. En effet, vu le nombre de campagnes et leur nombre important d'observations, un algorithme qui estime automatiquement le nombre de classes serait efficace. De ce fait, on a choisit l'algorithme Clues, qui existe dans le package R, qui est efficace en temps de traitement mais qui donne une solution locale. D'autres algorithmes existent comme ClusGap[8], mais il exige une allocation d'espace de mémoire trop importante pour le calcul de traitement, qu'on n'a pas pu l'exécuter vu la capacité limitée du matériel informatique dont on dispose.

[8] Tiré de http://stat.ethz.ch/R-manual/R-devel/library/cluster/html/clusGap.html

Pour beaucoup de campagnes, un nombre important de données nulles au niveau de la variable clic est constaté. Afin d'obtenir des relations significatives entre les différentes variables pendant la classification non supervisée, nous avons jugé pertinent de ne prendre en compte que les données dont les valeurs de la variable clic sont strictement positive.

Plusieurs tests ont été effectués sur différentes campagnes par l'algorithme Clues, un extrait des résultats obtenus est illustré dans le tableau suivant :

Tableau 4.4: Résultat de classification non supervisée par l'algorithme Clues

Campagne	Nombre observations	Nombre observations (clic>0)	Nombre mots-clés	Nombre de clusters	Indice Silhouette	Taille des clusters
001	132794	2942	2869	12	0.3555	158-176-207-328-167-282-492-335-277-191-134-195
002	359974	15919	14459	3	0.2865	5803-6528-3588
003	148830	12115	11246	4	0.2341	3809-3690-2596-2020
004	168810	13871	13336	2	0.4389	5080-8791
005	232768	18861	16772	2	0.4385	6733-12128
006	317640	25986	21374	2	0.4352	9212-16774
007	481598	30976	25230	2	0.3833	16532-14444
008	168810	13871	13336	2	0.4389	5080-8791
009	17993	2079	2044	2	0,3561	1288-791
010	15837	2389	2382	3	0,3518	601-1081-707
011	16830	4468	2984	2	0,5542	1200-3268
012	18573	5996	3156	2	0,412	3145-2851
013	795332	212488	75444	2	0.51	119251 - 93237
014	1195274	270083	79444	2	0.48	146497 - 123586

La connaissance au préalable d'un nombre de clusters nous permet d'explorer d'autres algorithmes nécessitant des traitements itératifs en fonction de la variation du paramètre d'entrée (nombre de classes) en un temps judicieux. En effet, on réduisant la variation de ce nombre aux alentours du résultat trouvé par l'algorithme Clues, ceci nous aide à réduire le temps de traitement du calcul et la capacité d'allocation de mémoire.

Nous avons constaté que la fréquence d'apparition du nombre de clusters égal à deux est la plus manifestée par l'algorithme Clues.

Des tests sur les combinaisons de campagnes également ont été effectués et nous ont donné approximativement le même résultat (voir tableau ci-dessous).

Tableau 4.5: Résultat de classification non supervisée des données annuelles de campagnes publicitaires par l'algorithme Clues

Campagne	Année	Nombre observations	Nombre observations (clic>0)	Nombre mots-clés	Nombre de clusters	Indice Silhouette	Taille des clusters
013	2010	3785	830	239	2	0,4738	536-294
014	2011	8475	1183	335	2	0,514	909-274
0025	2012	3359	556	209	2	0,3255	336-220
0026	2009	1506	755	303	2	0,3849	379-376
0026	2010	10534	4173	479	2	0,5028	1605-2568
0026	2011	11964	4616	475	2	0,519	2189-2427
0026	2012	9467	2413	704	6	0,4532	306-828-278-288-451-262
0027	2007	214	117	28	3	0,3416	44-53-20
0027	2008	7711	1003	29	2	0,4822	648-355
0027	2009	2771	729	95	2	0,3758	430-299
0027	2010	3321	874	91	2	0,3771	358-516
0027	2011	2575	494	45	8	0,3567	110-74-73-39-42-41-61-54
0027	2012	1150	300	91	5	0,3642	54-63-60-49-74
0028	2009	933	662	150	2	0,4061	233-429
0028	2010	2760	2164	271	2	0,461	1481-683
0028	2011	10150	5612	351	2	0,4796	4040-1572
0028	2012	8124	3419	317	2	0,3359	1235-2184
0029	2007	1802	1238	186	2	0.5271	859-379
0029	2008	3060	1869	386	3	0.3022	766-722-381
0029	2009	8806	4045	447	4	0.2434	1015-1747-609-674
0029	2010	14600	6051	910	2	0.4122	4359-1692
0029	2011	808	223	202	3	0.3007	89-82-52
0030	2010	3097	518	166	2	0,3126	347-171
0030	2011	6900	937	274	2	0,3961	501-436
0030	2012	2324	318	147	2	0,4598	105-213

Tel qu'expliqué à la section 3.3.4 l'estimation du nombre de classes par un tableau de classement (Tableau 3.5) nous donne un score sur les nombres décelés par les classifications. On constate que le nombre de classes égal à deux est largement dominant.

Méthodes appliquées

À la section 3.3.4, nous avons présenté quatre méthodes de classification non supervisées à appliquer. On a procédé à une série de tests pour chaque campagne mis en expérience tout en faisant varier les paramètres. L'utilité de chacune des méthodes est déterminée par l'indice de Silhouette. Rappelons que les algorithmes que nous avons mis en place se basent sur la distance euclidienne pour mesurer la similarité entre les individus.

On a utilisé successivement les algorithmes, l'un affinant les résultats des autres. Un modèle de résultats obtenus est illustré sous forme d'un tableau.

Tableau 4.6: Résultat de classification non supervisée par les algorithmes k-means et Clara

		Algorithme K-means		Algorithme Clara	
Nombre observation (clic >0)	**Nombre cluster**	**Indice Silhouette**	**Taille des clusters**	**Indice Silhouette**	**Taille des clusters**
15945	2	0,521	12050 - 3895	0,51	5259 - 10686
	3	0,45	3358 - 8004 - 4583	0,444	4408 - 6999 - 4538
Nombre observations total	4	0,464	4144 - 6579 - 3954 - 1268	0,56	3668 - 5765 - 2160 - 4352
44551	5	0,459	2575 - 4189 - 4997 - 3359 - 825	0,45	3303 - 3226 - 3499 - 1401 - 4516
	6	0,486	3004 - 3396 - 1818 - 2009 - 619 - 5099	0,491	2116 - 2598 - 4326 - 2017 - 2067 - 2821
	7	0,491	3323 - 4713 - 2101 - 1079 - 2464 - 676 - 1589	0,53	2179 - 3390 - 3410 - 1790 - 1361 - 3589 - 226
	8	0,455	761 - 2115 - 2028 - 3225 - 3224 - 1393 - 247 - 2952	0,543	1935 - 2712 - 4400 - 1290 - 1193 - 2270 - 1933 - 212
	9	0,481	878 - 580 - 354 - 3283 - 2086 - 3622 - 1166 - 1987 - 1989	0,484	1516 - 2081 - 1776 - 2360 - 1634 340 - 2608 - 1640 - 1990
	10	0,466	709 - 2079 - 161 - 2876 - 2684 - 559 - 2055 - 1340 - 1091 - 2391	0,474	1589 - 1963 - 2276 - 1306 - 372 1051 - 3206 - 2298 - 1799 - 85

Tableau 4.7: Résultat de classification non supervisée par les algorithmes Pam et C-means

		Algorithme Pam		Algorithme C-means	
Nombre observations (clic >0)	**Nombre cluster**	**Indice Silhouette**	**Taille des clusters**	**Indice Silhouette**	**Taille des clusters**
15945	2	0,48	5275 - 10670	0,51	11639 - 4306
	3	0,45	4436 - 7030 - 4479	0,45	7946 - 3340 - 4659
Nombre observations total	4	0,44	4042 - 2411 - 5236 - 4256	0,461	4098 - 6066 - 1588 - 4193
44551	5	0,48	3562 - 2117 - 5006 - 2109 - 3151	0,466	1168 - 2274 - 3820 - 3488 - 5195
	6	0,47	2106 - 3456 - 3903 - 2077 - 1252 - 3151	0,479	4413 - 870 - 3151 - 3563 - 1928 - 2020
	7	0,45	2269 - 3014 - 1519 - 969 - 3051 - 3005 - 2118	0,464	3491 - 2006 - 629 - 2188 - 3326 - 1380 - 2925
	8	0,47	2248 - 3028 - 1390 - 1308 - 880 - 2961 - 2083 - 2047	0,471	1245 - 2089 - 2218 - 2144 - 3288 - 1051 - 3301 - 609
	9	0,46	2269 - 3014 - 1334 - 899 - 712 3051 - 1744 - 1661 - 1261	0,462	1311 - 2962 - 2735 - 2155 - 2162 834 - 1241 - 465 - 2080
	10	0,45	1610 - 2026 - 1044 - 2559 - 885 660 - 2463 - 1744 - 1680 - 1274	0,465	1390 - 2594 - 1301 - 612 - 2845 1960 - 1939 - 2148 - 767 - 389

Une fois la classification non supervisée des mots-clés obtenue, on souhaite savoir quelles sont les variables responsables de tels regroupements. De ce fait, on a intégré dans notre algorithme la fonction « klaR » de R qui détermine les variables les plus influençables à la séparation des classes. Elle utilise la méthode d'analyse de discrimination linéaire.

Elle permet de détecter rapidement les principales variables qui différencient fortement les groupes. Ainsi les résultats obtenus révèlent la variable « position » qui discrimine fortement la séparation (voir l'exemple suivant : Tableau 4.8).

Tableau 4.8: Résultat de classification non supervisée selon Clues secondé par Clara et Pam

				Algorithme Clues				Algorithme Clara (nombre cluster = 2)		Algorithme Pam (nombre cluster = 2)	
Campagne	Nombre observations	Nombre observations (clic>0)	Nombre mots-clés	Nombre clusters	Indice Silhouette	Taille des clusters	Variables discriminantes	Indice Silhouette	Centroide Position	Indice Silhouette	Centroide Position
001	19309	934	263	2	0,4268	540-394	POSITION_MOY	0,5	1,773	0,5	1,738
002	14126	769	201	2	0,5469	440-329	POSITION_MOY	0,6	1,858	0,5	1,485
003	29384	1058	206	2	0,4971	545-513	POSITION_MOY	0,6	2,062	0,5	2,163
004	23219	447	201	2	0,5148	316-131	POSITION_MOY	0,5	2,208	0,5	2,279
005	32935	703	345	2	0,5123	428-275	POSITION_MOY	0,5	1,705	0,5	1,643
006	19489	474	250	3	0,4527	161-185-128		0,5	2,325	0,5	2,275
007	31948	490	250	2	0,5258	304-186	POSITION_MOY	0,6	2,301	0,5	2,275
008	24183	679	242	2	0,4490	357-322	POSITION_MOY	0,5	2,46	0,5	2,309
009	23939	923	380	2	0,5094	517-406	POSITION_MOY	0,6	1,742	0,5	1,784
010	32993	694	273	2	0,5253	370-324	POSITION_MOY	0,5	2,275	0,5	2,124
011	44944	1139	401	2	0,5609	823-316	POSITION_MOY	0,5	1,643	0,5	1,784
012	33957	926	294	2	0,5422	573-353	POSITION_MOY	0,6	2,124	0,5	2,275
013	65928	1397	420	3	0,415	527-457-413		0,6	1,725	0,5	1,889
014	135371	2385	448	2	0,4679	1160-1225	POSITION_MOY	0,6	1,882	0,5	2,062

D'après les résultats présentés dans ce tableau, l'algorithme Clara à une meilleure classification par rapport aux algorithmes Clues et Pam. Notons que les cases vides au niveau de la variable discriminante (colonne 8 du Tableau 4.8) indique qu'il n y a pas de variable sélectionnée pour cette séparation.

Résultats expérimentaux

Nous avons testé les différents algorithmes cités ci-dessus, on conclut que k-means et fuzzy c-means donnent pratiquement les mêmes résultats. Du point de vue du temps de traitement, k-means s'avère plus efficace mais le temps de calcul de son indice de silhouette s'avère trop long. Quant à l'algorithme Pam, qui nécessite encore plus de temps de traitement, son application reste fastidieuse.

Globalement, nous avons trouvé que l'algorithme k-means donne de meilleurs résultats selon l'indice de Silhouette.

Finalement, nous avons conclu que le nombre de classes représentatif de toute classification non supervisée de données de campagnes publicitaire est égal à deux.

4.4 Classification supervisée[9]

Tel que vu à la section précédente, le nombre de classe est égal à deux et la variable discriminante est la variable position.

Dans cette section, on procède au classement des données des mots-clés d'une campagne publicitaire par la méthode de classification supervisée en utilisant l'algorithme k-means dans le but de déterminer la valeur de la variable séparatrice « position ». k-means est un algorithme simple et efficace. Notre démarche est la suivante :

Pour un jeu de campagnes, on procède comme suit :

Étape 1 : appliquer l'algorithme k-means aux données d'une campagne en utilisant le paramètre de nombre de classes égal à 2.

Étape 2 : d'après le résultat, assurer que la première variable discriminante est celle de la variable position (sinon on passe à une autre campagne).

Étape 3: mesurer la qualité du classement par le calcul de l'indice de silhouette.

Étape 4 : déterminer l'étendue de chaque classe (calcul de l'intervalle des classes par les fonctions min et max).

Étape 5 : regrouper les résultats de toutes les campagnes en question dans un tableau en s'assurant que les colonnes du tableau des intervalles de classes soient bien mises en ordre d'une campagne à une autre.

[9] Le terme supervisé est employé dans le cadre où le nombre de classe est connu

Étape 6 : pour les classes 1, calculer la moyenne pondérée des valeurs max des intervalles de la classe en question par l'indice silhouette associé.

Refaire la même chose pour la classe 2, mais cette fois-ci en prenant les valeurs min des intervalles de la classe 2.

Étape 7 : calculer la moyenne des deux résultats obtenus à l'étape 6.

Le résultat final est la valeur discriminante de la variable position.

Par le Tableau 4.9, on présente un extrait de résultats de l'étape jusqu'à l'étape 5. Les lignes soulignées en jaune dans le tableau seront écartées (selon l'étape 2) lors du calcul des moyennes pondérées (l'étape 6).

La formule du calcul de la moyenne pondérée de la variable séparatrice position est :

$$PosMoy^{[1]} = \frac{\sum_j Sil_j * Vmax_j{}^{[1]}}{\sum_j Sil_j}$$

$$PosMoy^{[2]} = \frac{\sum_j Sil_j * Vmin_j{}^{[2]}}{\sum_j Sil_j}$$

où Sil_j est l'indice de silhouette calculé au niveau de la campagne *i*.

$Vmax_j{}^{[1]}$ est la valeur maximale de l'intervalle de la classe 1 (de la variable position) associé à la campagne j.

$Vmin_j{}^{[2]}$ est la valeur minimale de l'intervalle de la classe 2 (de la variable position) associé à la campagne j.

La valeur de la variable séparatrice « position » sera

$$position\ séparatrice = \frac{1}{2}(PosMoy^{[1]} + PosMoy^{[2]})$$

Tableau 4.9: Résultats de classification supervisée par l'algorithme k-means

	Nbre Observation (Clic >0)	Indice Silhouette	Taille des clusters	Variables Discriminantes	Perfermance de la discrimination	Centroides de la variable Descrimin. (C1 - C 2)	Classe 1 (Min - Max)	Classe2 (Min - Max)
1	20861	0.61	16184 - 4677	Position_moy - CPC_moy - Compte_mot	94.91% - 95.01% - 95.05%	1.71 - 4.19	1 - 2.96	2.94 - 12.97
2	24504	0.62	18927 - 5577	Position_moy - Compte_mot	95.82% - 95.86%	1.8 - 4.52	1 - 3.17	3.15 - 12.97
3	10663	0.57	7953 - 2710	Position_moy - CPC_moy - Compte_mot	95.61% - 95.61% - 95.61%	2.55 - 4.93	1 - 3.75	3.73 - 11
4	14417	0.57	10524 - 3893	Position_moy - CTR_moy - Somme_Impression	95.63% - 95.71% - 95.71%	2.59 - 4.99	1 - 3.79	3.79 - 11
5	2092	0.55	1240 - 852	Position_moy - CTR_moy - Somme_Coût	96.94% - 97.08% - 97.18%	2.75 - 5.18	1 - 3.96	3.97 - 9.96
6	4570	0.56	2827 - 1743	Position_moy - Somme_Impression - CTR_moy	94.9% - 95.48% - 95.9%	2.81 - 5.25	1 - 4.03	4.03 - 11
7	4030	0.61	2595 - 1435	Position_moy - Somme_Impression	97.1% - 97.12%	3.35 - 7.09	1 - 5.21	5.22 - 12.29
8	557	0.51	256 - 301	Position_moy - CTR_moy - Somme_Clic	98.92% - 99.1% - 99.46%	5.31 - 7.45	3 - 6.37	6.38 - 9.5
9	761	0.59	633 - 128	Position_moy - Compte_mot	97.11% - 97.37%	1.39 - 3.32	1 - 2.35	2.38 - 7
10	694	0.49	533 - 161	Position_moy - CPC_moy - Somme_Impression	92.22% - 93.37% - 94.09%	2.62 - 5.08	1 - 3.83	3.85 - 11.99
11	3595	0.64	2264 - 1331	Position_moy - Somme_Impression	98.92% - 98.94%	2.48 - 6.51	1 - 4.5	4.5 - 11
12	2968	0.52	1320 - 1648	Position_moy - CTR_moy	98.48% - 98.75%	6.04 - 8.95	1 - 7.51	7.48 - 11.93
13	730	0.38	557 - 173	Position_moy - Somme_Coût	93.97% - 93.97%	1.29 - 2.31	1 - 1.83	1.74 - 8.33
14	278	0.61	244 - 34	Position_moy - Somme_Clic	97.13% - 97.13%	1.71 - 4.25	1 - 2.96	3 - 6.35
15	25730	0.56	19130 - 6600	Position_moy - Compte_mot - Somme_Impression	94.8% - 94.82% - 94.82%	1.65 - 3.48	1 - 2.57	2.55 - 11
16	2451	0.56	1775 - 676	Position_moy - Compte_mot	95.1% - 95.19%	1.82 - 4.14	1 - 2.97	2.97 - 8.93
17	1271	0.25	774 - 497	CPC_moy - Position_moy - Compte_mot	82.85% - 88.91% - 94.57%	0.08 - 0.36	0 - 0.52	0 - 1
18	20338	0.55	15504 - 4834	Position_moy - CPC_moy - Compte_mot	94.76% - 94.85% - 94.88%	1.58 - 3.17	1 - 2.4	2.36 - 11.61
19	339	0.54	188 - 151	Position_moy - CTR_moy	98.82% - 99.71%	2.91 - 5.72	1 - 4.28	4.32 - 8.4
20	1897	0.55	885 - 1012	Position_moy - CTR_moy	99.47% - 99.52%	3.27 - 6.61	1 - 4.93	4.94 - 11
21	438	0.5	249 - 189	Position_moy - Compte_mot - CTR_moy	97.27% - 97.72% - 98.18%	3.07 - 5.66	1 - 4.37	4.37 - 10
22	1253	0.57	984 - 269	Position_moy - CPC_moy - Somme_Impression	96.25% - 96.49% - 96.49%	1.74 - 4.24	1 - 2.97	3 - 7.63
23	741	0.51	505 - 236	Position_moy - Compte_mot	96.09% - 96.22%	1.85 - 4.33	1 - 3.08	3.05 - 8.3
24	1414	0.6	1134 - 280	Position_moy - Compte_mot	97.81% - 97.95%	1.35 - 3.2	1 - 2.27	2.27 - 6.65
25	977	0.58	727 - 250	Position_moy - Compte_mot - Somme_Clic	95.5% - 95.8% - 96.01%	3.13 - 6.79	1 - 4.95	4.97 - 13
26	545	0.37	309 - 236	Position_moy - CTR_moy	96.89% - 97.07%	2.11 - 3.33	1.09 - 2.74	2.71 - 5.59
27	996	0.42	697 - 299	Position_moy - Somme_Clic - Somme_Impression	94.58% - 94.78% - 94.78%	1.96 - 3.64	1 - 2.89	2.76 - 7
28	1603	0.51	1000 - 603	Position_moy - Compte_mot - Somme_Impression	95.45% - 95.88% - 96.01%	3.09 - 5.6	1 - 4.34	4.35 - 11.07
29	3367	0.55	2165 - 1202	Position_moy - Somme_Impression	96.64% - 96.64%	2.85 - 5.38	1 - 4.11	4.12 - 9.65
30	1664	0.46	1013 - 651	Position_moy - Somme_Clic	98.08% - 98.38%	1.39 - 2.47	1 - 1.94	1.9 - 4.22
31	1370	0.67	1042 - 328	Position_moy - Somme_Impression - Compte_mot	97.66% - 97.74% - 97.74%	1.99 - 5.69	1 - 3.84	3.85 - 10.35
32	1217	0.55	914 - 303	Position_moy	94.82%	4.07 - 7.81	1 - 5.94	5.91 - 12.09
33	1130	0.39	296 - 834	Compte_mot - CTR_moy	97.88% - 97.96%	2.29 - 3.89	1 - 2.63	2.57 - 5
34	597	0.44	168 - 429	Position_moy	95.64%	1.33 - 2.81	1 - 2.1	1.94 - 6.07
35	23173	0.43	16965 - 6208	Position_moy - Somme_Impression	93.13% - 93.16%	1.84 - 3.3	1 - 2.69	2.54 - 12.25
36	287	0.23	69 - 218	CTR_moy	98.97%	0.42 - 0.95	0 - 0.65	0.69 - 1
37	9638	0.4	6263 - 3375	Position_moy - Compte_mot - CPC_moy	91.71% - 92.53% - 92.64%	1.81 - 3.24	1 - 2.69	2.49 - 12
38	1761	0.57	1344 - 417	Position_moy - Somme_Clic	97.33% - 97.67%	1.34 - 3.25	1 - 2.33	2.17 - 6.35
39	295	0.46	176 - 119	Position_moy - Compte_mot	96.61% - 97.29%	1.74 - 3.31	1 - 2.52	2.51 - 6.36

Tableau 4.9 : Résultats de classification supervisée par l'algorithme k-means (suite et fin)

	Nbre Observation (Clic >0)	Indice Silhouette	Taille des clusters	Variables Discriminantes	Perfermance de la discrimination	Centroides de la variable Descrimin. (C1 - C 2)	Classe 1 (Min - Max)	Classe2 (Min - Max)
40	903	0.42	722 - 181	Position_moy - CPC_moy - CTR_moy	93.79% - 94.68% - 94.9%	1.12 - 1.88	1 - 1.54	1.41 - 4.14
41	1127	0.45	440 - 687	Position_moy - Somme_Impression	95.92% - 96.36%	5.86 - 7.64	2.71 - 6.75	6.75 - 11.99
42	222	0.45	185 - 37	Position_moy - Somme_Impression - CPC_moy	95.06% - 96.88% - 97.77%	1.75 - 3.06	1 - 2.42	2.38 - 5.6
43	1670	0.5	1293 - 377	Position_moy - Compte_mot	92.81% - 92.99%	1.51 - 2.69	1 - 2.1	2.1 - 6
44	536	0.5	249 - 287	Position_moy - CTR_moy	98.14% - 98.51%	1.46 - 3.21	1 - 2.43	2.23 - 7.02
45	1806	0.59	1315 - 491	Position_moy - Somme_Impression	95.41% - 95.41%	2.49 - 5.31	1 - 3.9	3.91 - 10.5
46	163	0.5	110 - 53	Position_moy - Somme_Clic - Somme_Impression	95.77% - 97.57% - 97.57%	2.66 - 5.21	1.39 - 3.95	3.9 - 10.03
47	203327	0.49	127674 - 75653	Compte_mot - Position_moy	90.49% - 92.61%	2.93 - 1.14	1 - 4	1 - 4
48	248	0.49	152 - 96	Position_moy	98,00%	4.05 - 6.26	2 - 5.12	5.17 - 9.45
49	195	0.51	91 - 104	Position_moy - Somme_Impression	97.47% - 98%	2.38 - 4.57	1.47 - 3.52	3.37 - 9.77
50	1266	0.51	946 - 320	Position_moy - Somme_Impression	92.81% - 93.76%	2.82 - 5.66	1 - 4.24	4.23 - 12.48
51	954	0.5	516 - 438	Position_moy - Somme_Clic	98.43% - 98.63%	3.94 - 6.93	1 - 5.43	5.42 - 11.55
52	713	0.5	546 - 167	Position_moy - CPC_moy	95.23% - 95.66%	2.07 - 3.79	1 - 2.96	2.9 - 8
53	5168	0.66	3406 - 1762	Position_moy	98.63%	2.58 - 6.89	1 - 4.73	4.73 - 10.3
54	990	0.54	701 - 289	Position_moy - CPC_moy - Compte_mot	97.47% - 97.68% - 97.78%	1.34 - 3.09	1 - 2.21	2.2 - 6.17
55	1174	0.56	741 - 433	Position_moy - CTR_moy - Somme_Impression	97.1% - 97.27% - 97.27%	3.52 - 6.17	1 - 4.84	4.84 - 9.5
56	754	0.46	427 - 327	Position_moy - Somme_Impression	98.4% - 98.54%	2.95 - 5.04	1 - 4.01	3.99 - 7.92
57	965	0.51	734 - 231	Position_moy - CPC_moy - Somme_Impression	94.2% - 94.83% - 94.93%	1.78 - 3.59	1 - 2.7	2.68 - 7.95
58	2599	0.41	1547 - 1052	Position_moy - Compte_mot - Somme_Impression	93.61% - 96.92% - 96.96%	1.75 - 3.19	1 - 2.67	2.37 - 9
59	635	0.45	424 - 211	Position_moy - Somme_Impression - Compte_mot	96.85% - 97.16% - 97.16%	1.18 - 2.39	1 - 1.8	1.66 - 4.5
60	2200	0.47	1288 - 912	Position_moy - Somme_Impression	96.68% - 96.73%	2.73 - 4.9	1 - 3.8	3.81 - 9.36
61	1012	0.43	725 - 287	Position_moy - Compte_mot	95.75% - 95.95%	1.79 - 2.94	1 - 2.38	2.34 - 5.45
62	660	0.62	523 - 137	Position_moy - Compte_mot - Somme_Impression	96.21% - 96.21% - 96.36%	1.71 - 4.67	1 - 3.17	3.19 - 10.33
63	3141	0.49	1864 - 1277	Position_moy - CTR_moy - Somme_Impression	94.62% - 96.28% - 96.28%	2.73 - 5.04	1 - 3.88	3.89 - 11.67
64	1937	0.55	1661 - 276	Position_moy - Somme_Clic	96.18% - 96.44%	1.29 - 2.99	1 - 2.17	2.06 - 6.38
65	453	0.54	267 - 186	Position_moy - Somme_Impression	97.79% - 98.01%	2.45 - 5.19	1 - 3.81	3.84 - 8.4
66	1989	0.68	1752 - 237	Position_moy - Compte_mot - Somme_Clic	98.04% - 98.29% - 98.29%	1.19 - 2.92	1 - 2.07	2 - 6.2
67	14124	0.59	11197 - 2927	Position_moy - Compte_mot - CPC_moy	95.05% - 95.09% - 95.1%	1.62 - 3.57	1 - 2.61	2.57 - 11.49
68	183	0.49	102 - 81	Position_moy - Somme_Clic	98.89% - 99.44%	2.45 - 4.93	1.14 - 3.64	3.69 - 7.5
69	902	0.42	721 - 181	Position_moy - CPC_moy	93.8% - 94.79%	1.12 - 1.88	1 - 1.54	1.41 - 4.14
70	3277	0.52	2217 - 1060	Position_moy - Somme_Coût	95.88% - 95.88%	2 - 4.07	1 - 3.04	3.03 - 7.92
71	236	0.5	142 - 94	Position_moy - CPC_moy	95.33% - 95.72%	3.01 - 5.57	1 - 4.3	4.26 - 10.2
72	1172	0.53	862 - 310	Position_moy - CPC_moy - Compte_mot	95.05% - 95.73% - 95.99%	1.31 - 2.85	1 - 2.14	2.05 - 5.67
73	575	0.65	426 - 149	Position_moy - Somme_Impression	99.48% - 99.48%	1.41 - 3.98	1 - 2.68	2.7 - 6.35
74	509	0.5	409 - 100	Position_moy - CPC_moy	94.69% - 95.28%	1.41 - 2.84	1 - 2.14	2.11 - 5.63
75	716	0.56	574 - 142	Position_moy - Somme_Clic	96.93% - 96.93%	1.28 - 2.81	1 - 2.04	2.05 - 6.41
76	5587	0.42	4159 - 1428	Position_moy - Somme_Coût - Somme_Clic	91.69% - 91.77% - 91.8%	1.84 - 3.41	1 - 2.73	2.6 - 10
77	445	0.59	304 - 141	Position_moy - Somme_Impression - CTR_moy	97.09% - 97.54% - 97.99%	2.07 - 5.36	1 - 3.71	3.75 - 11
78	698	0.54	507 - 191	Position_moy	94.84%	1.94 - 4.16	1 - 3.03	3.05 - 9.38
79	1552	0.54	1140 - 412	Position_moy - Somme_Clic - CTR_moy	96.84% - 97.1% - 97.1%	1.38 - 3.18	1 - 2.29	2.27 - 8

Tableau 4.10: Estimation de la valeur du spectre de la séparation

Campagnes	Indice Silhouette	Classe 1 (Min - Max)	Valeur Max de la classe 1	Valeur Min de la classe 2	Classe2 (Min - Max)
1	0.61	1 - 2.96	2.96	2.94	2.94 - 12.97
2	0.62	1 - 3.17	3.17	3.15	3.15 - 12.97
3	0.57	1 - 3.75	3.75	3.73	3.73 - 11
4	0.57	1 - 3.79	3.79	3.79	3.79 - 11
5	0.55	1 - 3.96	3.96	3.97	3.97 - 9.96
6	0.56	1 - 4.03	4.03	4.03	4.03 - 11
7	0.61	1 - 5.21	5.21	5.22	5.22 - 12.29
8	0.51	3 - 6.37	6.37	6.38	6.38 - 9.5
9	0.59	1 - 2.35	2.35	2.38	2.38 - 7
10	0.49	1 - 3.83	3.83	3.85	3.85 - 11.99
11	0.64	1 - 4.5	4.5	4.5	4.5 - 11
12	0.52	1 - 7.51	7.51	7.48	7.48 - 11.93
13	0.38	1 - 1.83	1.83	1.74	1.74 - 8.33
14	0.61	1 - 2.96	2.96	3	3 - 6.35
16	0.56	1 - 2.57	2.57	2.55	2.55 - 11
17	0.56	1 - 2.97	2.97	2.97	2.97 - 8.93
19	0.55	1 - 2.4	2.4	2.36	2.36 - 11.61
20	0.54	1 - 4.28	4.28	4.32	4.32 - 8.4
...	...	...	...	...	...
...	...	...	...	...	...
60	0.55	1 - 2.17	2.17	2.06	2.06 - 6.38
61	0.54	1 - 3.81	3.81	3.84	3.84 - 8.4
62	0.68	1 - 2.07	2.07	2	2 - 6.2
63	0.59	1 - 2.61	2.61	2.57	2.57 - 11.49
64	0.49	1.14 - 3.64	3.64	3.69	3.69 - 7.5
65	0.42	1 - 1.54	1.54	1.41	1.41 - 4.14
66	0.52	1 - 3.04	3.04	3.03	3.03 - 7.92
67	0.5	1 - 4.3	4.3	4.26	4.26 - 10.2
68	0.53	1 - 2.14	2.14	2.05	2.05 - 5.67
69	0.65	1 - 2.68	2.68	2.7	2.7 - 6.35
70	0.5	1 - 2.14	2.14	2.11	2.11 - 5.63
71	0.56	1 - 2.04	2.04	2.05	2.05 - 6.41
72	0.42	1 - 2.73	2.73	2.6	2.6 - 10
73	0.59	1 - 3.71	3.71	3.75	3.75 - 11
74	0.54	1 - 3.03	3.03	3.05	3.05 - 9.38
75	0.54	1 - 2.29	2.29	2.27	2.27 - 8
	Moyenne pondérée :		3.41	3.38	
	Moyenne des résultat		**3.4**		

Interprétation des résultats

L'analyse des résultats montre que la position à valeur 3,4 est bien un spectre de séparation. Il est surprenant que tous les tests effectués sur plusieurs campagnes aient révélé presque la même valeur de la position séparatrice (un écart $\pm$ 0.3 est observé). Rappelons que les analyses ont été faites sans tenir compte de la distinction entre les positions Premium et les positions standards. Par contre, les résultats permettent de conclure qu'il y a bien une distinction entre ces deux types de position.

4.5 Fonctions génériques et validation des paramètres

Fonctions génériques

Nous nous appuyons sur l'étude existante d'analyse de données (Quinn, 2001, p. 85). Nous avons proposé une approche permettant d'ajuster les paramètres existants clics et CPC des fonctions génériques pour chaque classe.

Tel qu'expliqué dans la section 3.3.6, face à l'hypothèse que tous les mots-clés possèdent sensiblement les mêmes taux de décroissance des fonctions de clics et de CPC qui sont caractérisées par les équations (2) et (3), nous allons extraire deux ensembles de mots-clés. Ces deux ensembles vont nous permettent de constituer des modèles de fonctions de prédiction de qualité associent aux clics et aux CPC selon les critères sus cités (page 47).

Illustrons un exemple de fonctions génériques des taux de décroissance de clics et de CPC relatif à la position d'un mot-clé donné extrait à partir de l'ensemble filtré de notre base de données.

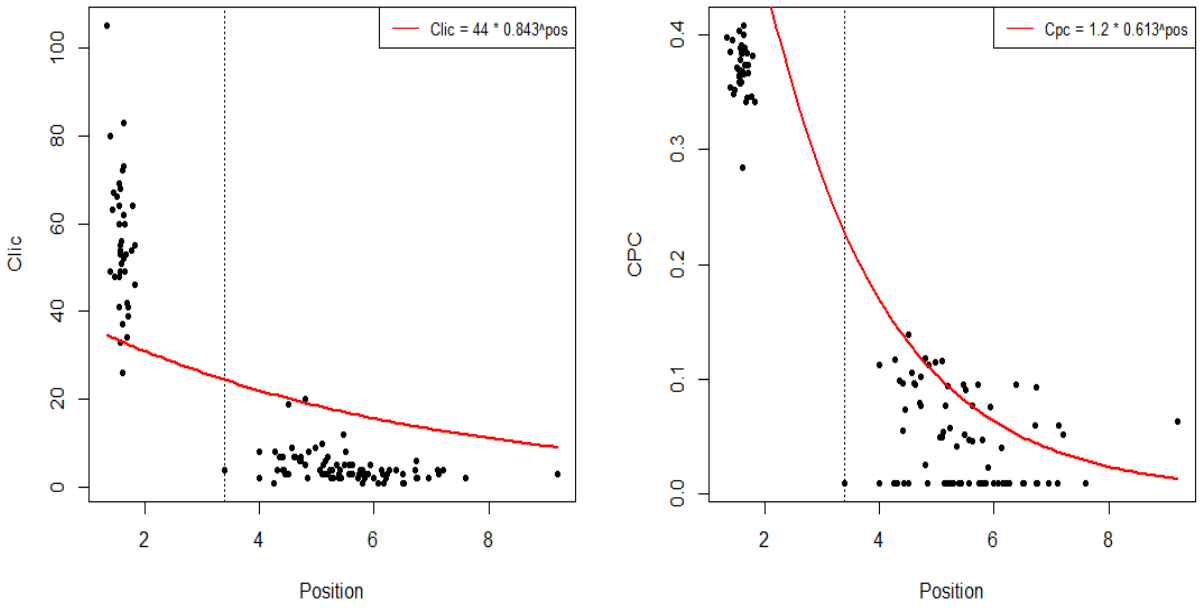


Figure 4.6: Représentations graphiques respectives des fonctions génériques du clic et du CPC d'un mot-clé par rapport à la position

Selon notre méthode de classification, un seul critère sera modifié qui dépend de la position de la valeur séparatrice « 3,4 ». Le critère est le suivant :

- valeur de position minimale ≤ 2 (on s'assure qu'au moins une observation se situe entre 1 et 2 pour qu'il y ait l'inclusion de la valeur séparatrice)

Dans ce cas, on doit déterminer les paramètres $c_{clic}^{[1]}$ et $c_{cpc}^{[1]}$; $k_{clic}^{[1]}$ et $k_{cpc}^{[1]}$ pour la classe 1 et $c_{clic}^{[2]}$ et $c_{cpc}^{[2]}$; $k_{clic}^{[2]}$ et $k_{cpc}^{[2]}$ pour la classe 2.

Les paramètres $c_{clic}^{[1]}$, $c_{cpc}^{[1]}$, $c_{clic}^{[2]}$ et $c_{cpc}^{[2]}$ sont estimés par la moyenne des valeurs observées du taux de décroissance des clics et des CPC calculés sur l'ensemble des mots-clés obtenus par les critères exigés.

Pour y arriver, les formules (2) et (3) seront linéarisées par la fonction logarithme suivante :

$$\ln(clic) = \ln(k_{clic}) + \ln(c_{clic}) * pos$$

$$\ln(cpc) = \ln(k_{cpc}) + \ln(c_{cpc}) * pos.$$

Ensuite, on effectue des régressions linéaires avec les observations $(pos, \ln(clic))$ et $(pos, \ln(cpc))$ pour chaque mot-clé extrait selon les critères afin de déterminer les paramètres globaux.

On subdivise l'ensemble des mots-clés obtenus selon les critères sus cités, par la valeur séparatrice en deux sous-ensembles pour déterminer les paramètres globaux de chaque classe.

Finalement, les paramètres $c_{clic}^{[1]}$, $c_{cpc}^{[1]}$, $c_{clic}^{[2]}$ et $c_{cpc}^{[2]}$ sont estimés par la moyenne des valeurs observées des paramètres de la régression linéaire $(\ln(clic))$ de chaque mot-clé, transformées par la fonction exponentielle.

Les résultats des moyennes ainsi obtenues pour chaque classe sont :

$$c_{clic}^{[1]} = 0{,}8442 \quad et \quad c_{cpc}^{[1]} = 0{,}7634$$

$$c_{clic}^{[2]} = 0{,}7036 \quad et \quad c_{cpc}^{[2]} = 0{,}7092$$

Concernant les paramètres constants propres à chaque mot-clé $k_{clic}^{[1]}$, $k_{cpc}^{[1]}$, $k_{clic}^{[2]}$ et $k_{cpc}^{[2]}$ caractérisant l'échelle de grandeur de chaque courbe, ils sont calculés selon les formules des équations (4) et (5) tout en tenant compte de la séparation des observations du mot-clé par rapport aux classes.

Illustrons par un graphique les fonctions génériques des variables clic et CPC associées au mot-clé d'une campagne publicitaire donnée en fonction de la variable position, dont le comportement du mot est séparé en deux classes. La fonction génératrice associée à la classe 1 est représentée par la couleur bleu et celle de la classe 2, par la couleur verte.

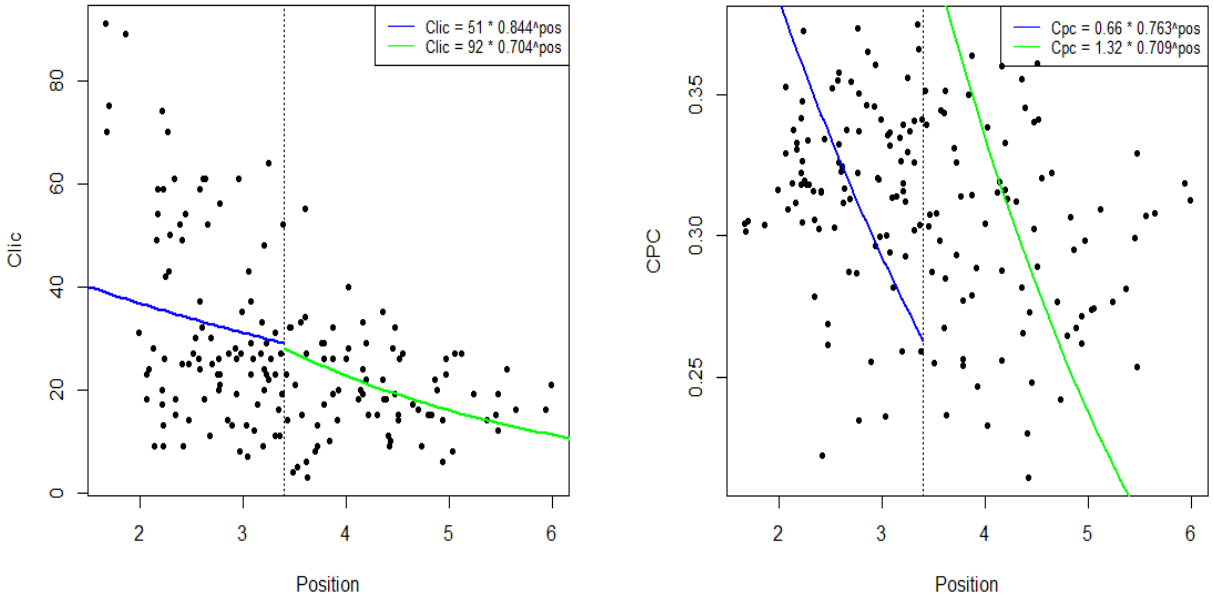


Figure 4.7: Représentation des fonctions génériques clic et CPC d'un mot-clé selon la classification

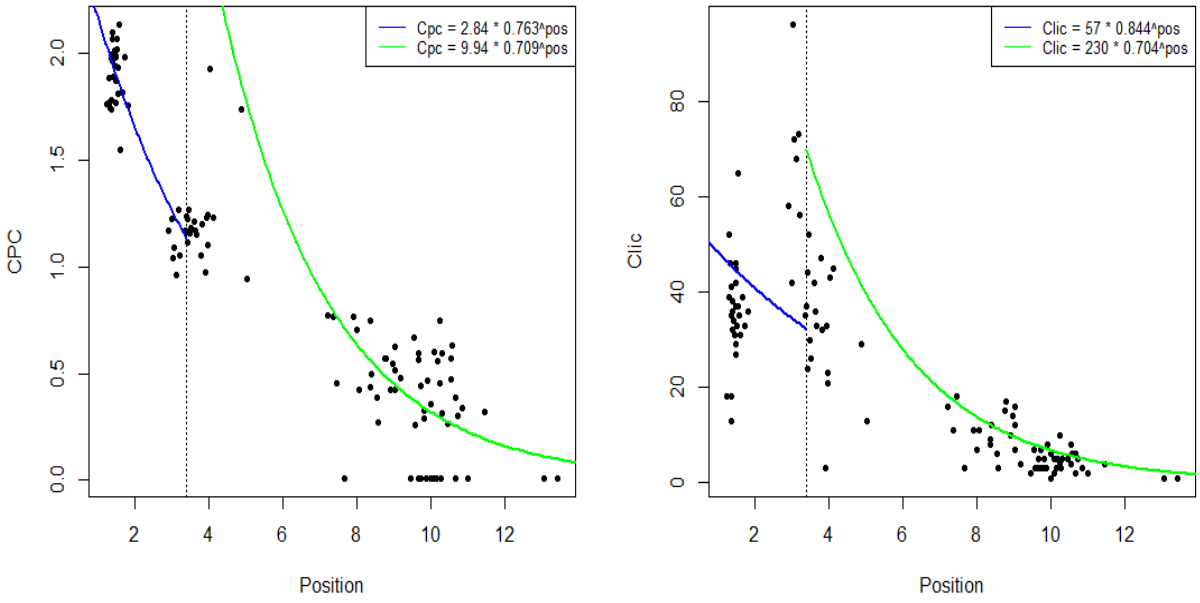


Figure 4.8: Représentation graphique des fonctions génériques d'un autre mot-clé

Validation des paramètres

Afin de s'assurer que notre approche peut s'étendre sur l'ensemble des mots-clés de notre base de données, une validation de nos paramètres s'avère nécessaire. L'application de cette analyse de validation est déjà expliquée dans la section 3.3.7.

Le détail des résultats de notre expérience est représenté sous forme de tableau (Tableau 4.11 pour le variable clic et Tableau 4.12 pour la variable CPC) de calcul d'erreurs de prédictions de la variable clic et CPC pour les 30 prochains jours, basées sur les paramètres des fonctions génériques qui sont établis sur une période de 120 jours de données historiques.

Tableau 4.11: Calcul d'erreur de prédictions de la variable clic pour chaque mot-clé d'une campagne donnée

Mot-clé	Nb. obs. Classe 1	$k_{clic}^{[1]}$	$MAE^{[1]}$	Nb. obs. Classe 2	$k_{clic}^{[2]}$	$MAE^{[2]}$	Nb. obs. total	MAE
0001	27	288,0013254	93,996155	3	345,506362	82,10673183	30	176,10289
0002	22	99,11709144	28,504361	8	436,575918	38,95449537	30	67,458856
0003	0	2,745299314	NaN	30	946,768624	62,28974781	30	62,289748
0004	4	190,550608	30,879814	26	305,779567	30,73691731	30	61,616732
0005	30	6,576836419	61,484556	0	2,47707513	NaN	30	61,484556
0006	27	19,75421744	36,733528	3	45,4508293	24,33550127	30	61,069029
0007	30	221,8549737	58,420302	0	347,500075	NaN	30	58,420302
0008	3	63,08969674	33,492995	27	94,8120899	20,08900965	30	53,582005
0009	10	48,99138804	22,2793	20	75,8759897	28,85872003	30	51,13802
0010	27	13,12059738	6,97066	3	151,660547	41,40076694	30	48,371427
0011	7	32,56259459	13,868486	23	171,565198	34,31039374	30	48,17888
0012	4	80,45498095	37,772816	26	79,4359894	8,083098306	30	45,855914
0013	30	115,7021019	38,304983	0	93,3630049	NaN	30	38,304983
0014	23	61,12982241	20,447768	7	29,8661012	17,57589379	30	38,023662
0015	24	2,106328597	27,770493	6	4,34133	7,507286739	30	35,27778
...	...	...	...	...	...	...	...	...
2536	9	0,021839089	0,0145712	10	0,21211509	0,1348115	19	0,1493827
2537	2	0,171876905	0,1034031	5	0,38726391	0,044941434	7	0,1483445
2538	9	0,081007622	0,0482911	16	0,43921117	0,091165374	25	0,1394564
2539	7	0,060897991	0,0396662	23	0,50537347	0,096280172	30	0,1359464
2540	25	0,047448258	0,0697065	5	0,28397996	0,065001402	30	0,1347079
2541	10	0,148185944	0,0994656	14	0,13734029	0,026813385	24	0,126279
2542	11	0,026766536	0,1039169	19	0,09098811	0,019655288	30	0,1235721
2543	1	0,186578298	0,1112068	25	0,07662721	0,009625971	26	0,1208327
2544	0	0,551388995	NaN	30	0,503631	0,087091681	30	0,0870917
2545	24	0,059095108	0,0420331	6	0,21352037	0,044993499	30	0,0870266
2546	16	0,052394132	0,039163	2	0,30297621	0,043355028	18	0,082518
2547	0	0,51003437	NaN	30	0,72929179	0,067959425	30	0,0679594
2548	5	0,073190292	0,0598609	0	1,21955578	NaN	5	0,0598609
2549	2	0,047443719	0,0338108	2	0,13123928	0,021500148	4	0,055311
2550	0	0,055631915	NaN	16	0,2792692	0,02620382	16	0,0262038
							$EG_{clic} =$	**2,86**

Tableau 4.12: Calcul d'erreur de prédictions de la variable CPC pour chaque mot-clé d'une campagne donnée

Mot-clé	Nb. obs. Classe 1	$k_{cpc}^{[1]}$	$MAE^{[1]}$	Nb. obs. Classe 2	$k_{cpc}^{[2]}$	$MAE^{[2]}$	Nb. obs. total	MAE
0001	27	1,300057942	0,3026372	3	2,13449016	0,209056035	30	0,5116932
0002	0	1,791640973	NaN	29	4,97103821	0,293359732	29	0,2933597
0003	30	1,549674157	0,06861	0	4,93727256	NaN	30	0,06861
0004	8	0,647378577	0,395252	22	1,02546185	0,429540205	30	0,8247922
0005	30	1,517104816	0,031213	0	4,68332452	NaN	30	0,031213
0006	0	1,857713937	NaN	25	4,38572974	0,152825029	25	0,152825
0007	16	1,061452304	0,537191	14	2,7210035	0,4666524	30	1,0038434
0008	27	1,961580458	0,2835232	3	3,27694875	0,189700927	30	0,4732241
0009	28	1,44498	0,2469784	2	2,97884695	0,151436782	30	0,3984152
0010	30	1,790302704	0,1059796	0	2,83843287	NaN	30	0,1059796
0011	30	1,857095214	0,1126432	0	4,09127489	NaN	30	0,1126432
0012	0	1,752956083	NaN	30	3,4059601	0,099377981	30	0,099378
0013	1	0,961267066	0,4061989	29	0,20201464	0,15281591	30	0,5590148
0014	18	0,595428963	0,4705574	2	0,35995445	0,105040192	20	0,5755976
0015	9	0,671585726	0,4276877	20	1,93537964	0,439228457	29	0,8669162
...	...	...	...	...	...	...	...	...
2536	7	1,04851827	0,1941479	23	1,16125613	0,073499641	30	0,2676475
2537	1	0,94978285	0,1574519	29	1,2112881	0,056847445	30	0,2142994
2538	30	2,239250646	0,156174	0	3,33754782	NaN	30	0,156174
2539	30	1,540995519	0,1146813	0	4,13205093	NaN	30	0,1146813
2540	30	4,722399544	2,072445	0	10,2913591	NaN	30	2,072445
2541	15	11,13052001	1,202718	15	21,330913	2,104300984	30	3,307019
2542	14	9,040654056	4,4889069	16	21,7933171	4,359376055	30	8,848283
2543	27	2,809686369	1,7415857	1	6,26498898	1,881610041	28	3,6231957
2544	17	4,439331534	2,6632263	13	10,7005071	2,796136888	30	5,4593631
2545	0	1,445748698	NaN	30	4,67567703	0,848180686	30	0,8481807
2546	27	3,161882699	1,5424026	3	2,38516245	2,776414076	30	4,3188167
2547	0	3,254034932	NaN	30	5,51094227	0,585190121	30	0,5851901
2548	30	2,828045484	0,1520522	0	6,19254943	NaN	30	0,1520522
2549	29	1,422205057	1,1193599	1	4,40307982	1,113624119	30	2,2329841
2550	12	4,201533426	1,719797	18	8,48489198	1,370535538	30	3,0903325
							$EG_{cpc} =$	**0,25**

Les marges d'erreurs associées à la fonction de prédiction clic (Tableau 4.11) sont relativement élevées d'un mot-clé à un autre. Mais au niveau d'une campagne où la marge d'erreur globale est estimée selon l'équation (7), le résultat est acceptable. L'interprétation de ce phénomène s'avère difficile tant la gestion des clics dépend de la complexité du comportement des internautes. Par contre, la marge d'erreur globale associées à la fonction de prédiction CPC est pratiquement faible (calculée selon l'équation (8)). En effet, les coûts des mots-clés sont gérés au niveau des agences. De ce fait, on peut interpréter qu'il y a une meilleure gestion de budget à ce niveau.

Afin de valider nos paramètres ainsi que la période des données historiques selon la méthode décrite à la Figure 3.7, les Tableaux 4.13 et 4.14 présentent les résultats globaux obtenus de la méthode.

Tableau 4.13: Résultats globaux pour l'estimation d'erreur de prédiction de la fonction clic

	Valider sur	Agence	Nombre mots-clés	$EG_{clic}^{[1]}$	$EG_{clic}^{[2]}$	EG_{clic}
parametre basé sur 30j	**mois 1**	A	3141	2,382	2,341	2,803
	mois 2			2,329	3,279	2,728
parametre basé sur 60j	**mois 1**			2,394	2,347	2,799
	mois 2			2,338	3,285	2,733
parametre basé sur 90j	**mois 1**			2,197	2,349	2,794
	mois 2			2,341	3,284	2,730
parametre basé sur 120j	**mois 1**			2,404	2,349	2,791
	mois 2			2,348	3,284	2,732

Tableau 4.14: Résultats globaux pour l'estimation d'erreur de prédiction de la fonction CPC

	Valider sur	Agence	Nombre mots-clés	$EG_{cpc}^{[1]}$	$EG_{cpc}^{[2]}$	EG_{cpc}
parametre basé sur 30j	**mois 1**	A	3141	0,433	0,411	0,891
	mois 2			0,392	0,366	0,834
parametre basé sur 60j	**mois 1**			0,439	0,412	0,879
	mois 2			0,393	0,364	0,831
parametre basé sur 90j	**mois 1**			0,441	0,408	0,869
	mois 2			0,394	0,360	0,825
parametre basé sur 120j	**mois 1**			0,447	0,407	0,866
	mois 2			0,399	0,359	0,828

On constate que le nombre important de données historiques n'influe pas trop sur la détermination des paramètres constants propres au mot-clé. Ainsi, une période récente de 30 jours de données historiques contribuera au mieux au calcul des paramètres constants.

4.6 Comparaison des fonctions génériques

Par une illustration graphique, représentons les fonctions génériques précédentes et actuelles afin d'avoir une idée sur la comparaison des résultats.

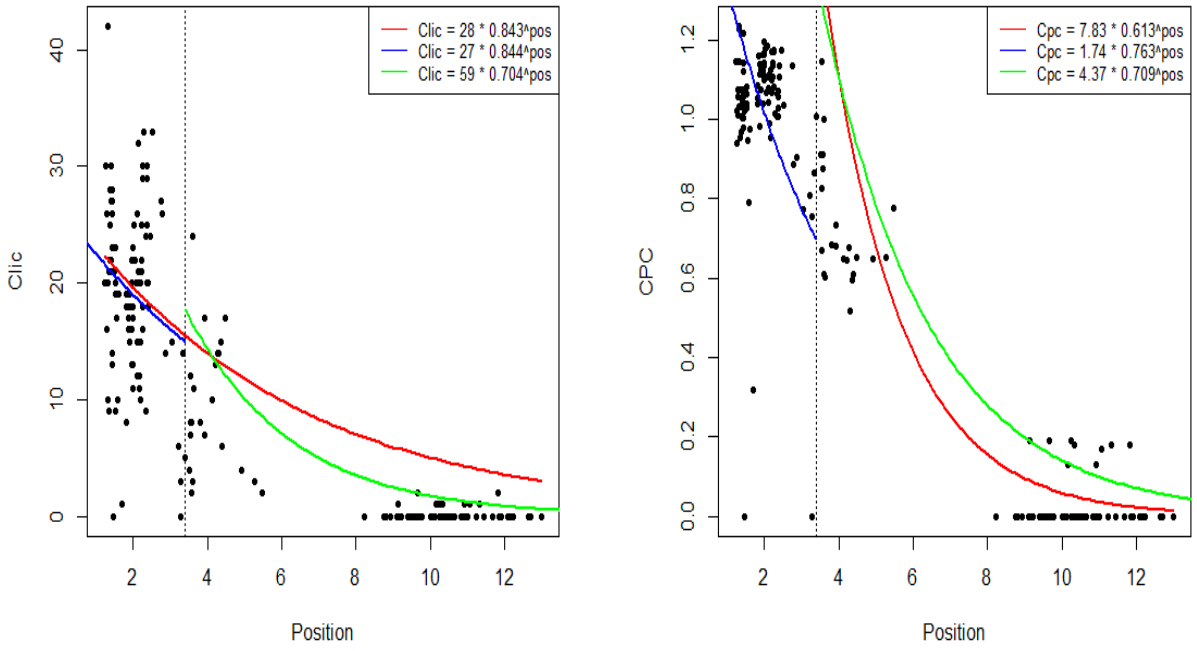


Figure 4.9: Exemple de graphique des fonctions génériques précédente (rouge) et actuelle (bleu et vert) pour la prédiction des clics et de CPC pour un mot-clé

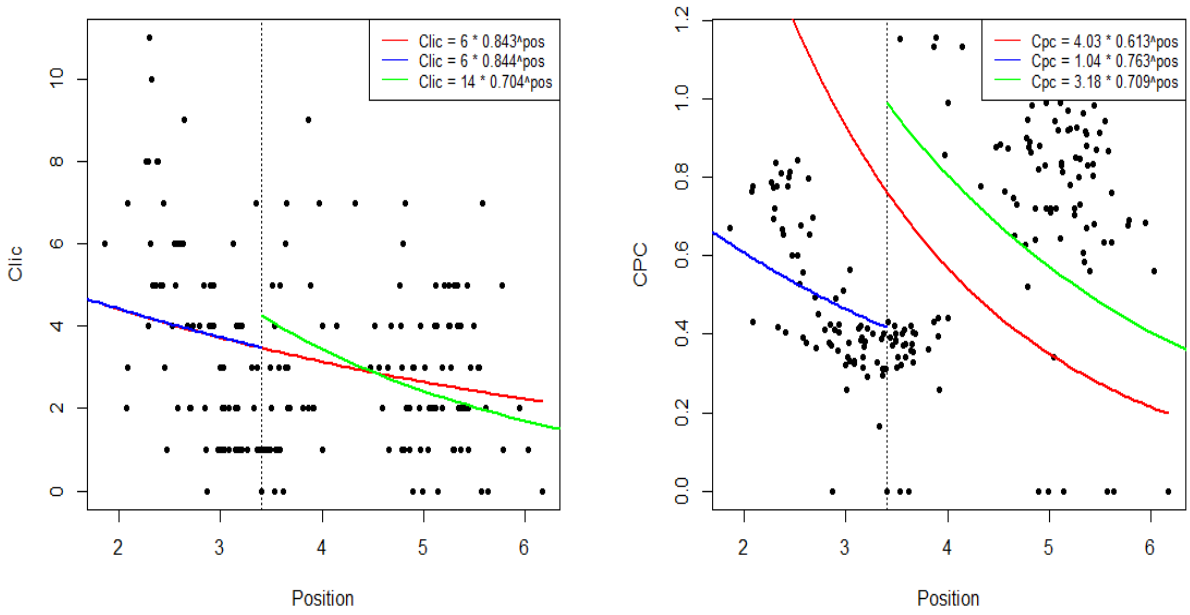


Figure 4.10: Autre exemple de comparaison de fonctions génériques

Les tableaux suivants nous illustrent les résultats analytiques selon la méthode décrite dans la section 3.3.8. Les valeurs des colonnes 3, 4, 6 et 7 des deux tableaux sont calculées selon l'équation (6) et les valeurs de la dernière colonne sont déterminées par l'équation (9).

Les résultats finaux sont les moyennes pondérées des rapports d'erreurs de prédictions entre les fonctions génériques améliorées et celle de la fonction générique précédente, dont le calcul est déterminé respectivement par les équations (10) et (11) associées aux fonctions clic et CPC.

Tableau 4.15: Résultat de comparaison des erreurs de prédiction de la fonction clic

Mot-clé	Nbre. obs. classe1	MAE 1 (ancienne)	MAE 1 (existante)	Nbre. obs. classe2	MAE 2 (ancienne)	MAE 2 (existante)	Nb. Obs. total	RCM
001	28	0,67981407	0,66326749	2	0,93103069	0,91428642	30	0,9793333
002	9	1,23560562	1,01726593	21	1,0506657	1,06291754	30	0,909858533
003	26	6,2016312	6,2892157	4	8,59013534	6,00945205	30	0,831453614
004	1	0,99809489	0,66674704	29	0,92491564	0,92896429	30	0,829798537
005	2	1,77384459	1	18	1,13080193	1,10202057	20	0,723675172
006	14	0,80934903	0,87372513	16	0,49624976	0,37617875	30	0,957341486
007	20	1,36838299	1,19652731	10	1,67553327	1,63985411	30	0,931819793
008	24	1,20962446	1,19405295	6	0,60928765	0,56344437	30	0,966235425
009	6	0,39199886	0,4188592	24	0,35154193	0,34285989	30	1,024448282
010	28	1,32086407	1,32641928	2	0,50045515	0,46292927	30	0,982446426
011	12	2,82278423	2,9306932	18	3,59208103	3,0619649	30	0,934183005
012	23	5,36992558	5,82523107	7	3,13966137	0,45584946	30	0,738118144
013	6	3,76500837	3,32939584	24	2,08775019	2,21342608	30	0,947044349
014	3	3,21408217	1,86096031	27	3,21705296	3,61543293	30	0,851543798
015	1	14,559655	13,7984862	29	6,83018036	6,93172379	30	0,96916174
016	23	7,51502476	7,57498611	7	9,42420349	8,33938476	30	0,939497988
017	27	1,40969706	1,43287895	2	1,64743226	0,50007591	29	0,632277755
018	1	1,82000379	2,07348616	29	12,6080471	12,4008231	30	1,003206139
019	9	1,07843568	0,94803169	21	1,91056502	1,90795277	30	0,95549809
020	7	2,62831227	2	23	1,88992765	2,07918671	30	0,902826494
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
2570	1	0,52018561	0,33333527	29	0,62339813	0,62567645	30	0,8386021
2571	19	0,96802805	0,95286932	11	0,49512543	0,44187647	30	0,9532464
2572	23	6,54329428	6,77169133	7	7,0030871	4,57142857	30	0,837354241
2573	5	4,29086177	4,57371468	21	13,9974347	6,83075681	26	0,623593974
2574	19	1,13639411	1,10385994	11	1,03092781	0,92059395	30	0,934080847
2575	1	1,86214059	1,83186343	29	0,11736554	0,11076723	30	0,981371382
2576	29	1,1274979	1,16874196	1	0,52630491	0,49980225	30	1,008913637
2577	13	6,09569189	5,445391	17	7,32736901	6,57232307	30	0,89530355
2578	18	3,65293234	4,21594049	12	3,61367973	2,19262335	30	0,881919082
2579	11	0,68539996	0,65166959	19	1,08142561	1,15761014	30	1,024028496
2580	26	3,5116272	3,58115303	4	4,75311362	2,17565375	30	0,696550189
2581	15	0,48388866	0,48758044	4	0,41306794	0,36355971	19	0,948920111
2582	26	1,53719542	1,71453942	4	1,65370357	1,23466842	30	0,924256097
2583	13	3,244452	3,43958789	17	3,43795813	3,10174257	30	0,978887906
2584	23	0,77968664	0,77691751	7	0,49163828	0,26158719	30	0,816868047
2585	23	1,13190182	1,08588415	7	0,83360525	0,46408667	30	0,788585729
2586	27	1,88434359	1,89060265	3	1,99987771	1,96805016	30	0,993417342
2587	15	0,44888303	0,50576324	14	0,19474745	0,07344279	29	0,899904589
2588	9	0,4855295	0,48561237	21	0,58026227	0,5743008	30	0,994484294
2589	15	0,48241282	0,29345036	15	0,39273485	0,44703146	30	0,846122142
2590	1	5,16454197	5,9995317	29	4,01056066	4,07996609	30	1,098570577
2591	4	6,75224209	4,75	26	5,73208339	5,85529457	30	0,849488792
2592	7	16,4540481	8,99863979	23	1,65573974	0,64977605	30	0,532773543
							ECGclic =	***0,905***

Tableau 4.16: Résultat de comparaison des erreurs de prédiction de la fonction CPC

Mot-clé	Nbre. obs. classe1	MAE 1 (ancienne)	MAE 1 (existante)	Nbre. obs. classe2	MAE 2 (ancienne)	MAE 2 (existante)	Nb. Obs. total	RCM
001	28	0,07567967	0,07555324	2	0,05044236	0,05072542	30	1,00124196
002	9	0,01844665	0,02677167	21	0,05781208	0,05828268	30	1,11533925
003	26	0,08709145	0,047079	4	0,12504127	0,08112838	30	0,60437342
004	1	0,10929496	0,09669622	29	0,06802432	0,06690699	30	0,92264763
005	2	0,08375262	0,03066667	18	0,05496594	0,05260606	20	0,60029981
006	14	0,08469678	0,08424071	16	0,06699971	0,06271578	30	0,96875339
007	20	0,03718546	0,02859222	10	0,05593783	0,05680864	30	0,917073
008	24	0,08967088	0,09081073	6	0,07702284	0,07254547	30	0,97997815
009	6	0,04201625	0,04330508	24	0,07935641	0,07858368	30	1,00425221
010	28	0,04171719	0,01447301	2	0,05469712	0,05167392	30	0,68606961
011	12	0,02606602	0,01126125	18	0,01745209	0,01725329	30	0,65523368
012	23	0,07597367	0,0751893	7	0,06715014	0,06447496	30	0,97582824
013	6	0,06797082	0,07101631	24	0,05316485	0,04970247	30	0,99655843
014	3	0,00259867	0,00157539	27	0,00614547	0,00498091	30	0,74979237
015	1	0,01213511	0,00459482	29	0,00265288	0,00474385	30	0,63150392
016	23	0,20116159	0,34804443	7	0,36803583	0,38321855	30	1,28472646
017	27	0,0726815	0,06446248	2	0,05336213	0,05368235	29	0,9373328
018	1	0,04281863	0,07719348	29	0,06177512	0,06162975	30	1,32726123
019	9	0,01598732	0,01299798	21	0,03098854	0,01640912	30	0,62600453
020	7	0,0461971	0,05235549	23	0,09381745	0,05313204	30	0,75340408
...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...
2570	1	0,11882122	0,12073107	29	0,07386739	0,07743681	30	1,02843587
2571	19	0,04052219	0,06648335	11	0,0676377	0,0668613	30	1,23284753
2572	23	0,08233475	0,04737661	7	0,05149664	0,05146383	30	0,73854453
2573	5	0,11651142	0,11643647	21	0,0844587	0,0971875	26	1,06296386
2574	19	0,40140204	0,2451875	11	0,04160086	0,0356322	30	0,63390037
2575	1	0,0524442	0,0532539	29	0,05427947	0,05560731	30	1,02002877
2576	29	0,01165861	0,007894	1	0,08002432	0,06863227	30	0,83468399
2577	13	0,05394834	0,05308852	17	0,06290354	0,05495183	30	0,92459232
2578	18	0,01769488	0,01274741	12	0,05287386	0,01495474	30	0,39255555
2579	11	0,14246914	0,01263882	19	0,0346666	0,02843155	30	0,23185819
2580	26	0,04544436	0,04752437	4	0,06319233	0,06058922	30	0,99518487
2581	15	0,07258585	0,0694712	4	0,07322009	0,07458009	19	0,98796591
2582	26	0,04341566	0,03884144	4	0,05835405	0,05841065	30	0,95560931
2583	13	0,07091166	0,07062822	17	0,02855318	0,06273684	30	1,34082616
2584	23	0,06299333	0,03383581	7	0,05517692	0,06114652	30	0,80377539
2585	23	0,03481829	0,02257103	7	0,05840443	0,03348362	30	0,60129819
2586	27	0,04797991	0,0510118	3	0,06944541	0,07171228	30	1,04512444
2587	15	10,6821076	7,51322015	14	2,90558418	2,75314273	29	0,75556342
2588	9	0,06490429	0,056	21	0,04459698	0,04269372	30	0,90130205
2589	15	0,02489105	0,02249865	15	0,08731845	0,026875	30	0,44001309
2590	1	0,07161074	0,02916667	29	0,05050116	0,04364273	30	0,59625143
2591	4	0,20431892	0,17042821	26	0,48596289	0,11982143	30	0,42047992
2592	7	0,04523307	0,04118083	23	0,04020721	0,04650341	30	1,02626351
							ECGcpc =	***0,874***

D'après les tests effectués, on constate que les fonctions génériques que nous avons raffinées demeurent plus efficaces que celles utilisées précédemment. Un ajustement de 10% a été observé.

Finalement, on termine notre travail par une conclusion présentée au chapitre suivant.

CHAPITRE 5 : CONCLUSION

Les sites Web de nos jours sont les outils de marketing les plus efficaces à influencer les gens qui y ont généralement accès grâce aux moteurs de recherche. Les annonces des sites les plus visualisés sont positionnées sur la première page des résultats des moteurs de recherche. Les formules de positionnement restent confidentielles, propres aux entreprises des moteurs de recherche. Ces entreprises ne dévoilent que les données statistiques relatives au nombre incommensurable de mots-clés associés aux annonces textuelles basées sur les moteurs de recherche.

Gérer un tel nombre de mots-clés exige certainement une catégorisation de ces mots. La question à poser est de savoir sur quelle information se base cette classification. Comme l'analyse des mots-clés de Google est confidentielle, il y a eu plusieurs recherches dans ce domaine. La plupart des recherches sont basées sur la sémantique des requêtes afin de suggérer les mots-clés adéquats afin d'obtenir les premières positions sur le moteur de recherche Google. Dans notre étude, nous optons pour l'analyse des mots-clés à partir de leurs données statistiques sur une large période.

Notre objectif se focalise sur le clustering des données par différents algorithmes automatiques de classification. Nous identifions les classes selon l'utilisation d'une distance euclidienne entre les mots-clés par leurs caractéristiques sur une base de données historique. Le taux de convergence des résultats a révélé deux modèles (classes) de connaissances pour l'ensemble des campagnes. Il est d'ailleurs rare en pratique qu'un problème de classification dépend d’une seule variable.

Les tests effectués montrent que la variable séparatrice significative est la variable position dont une valeur fait la distinction entre les positions premiums et les positions standards. Nous déduisons qu'il existe uniquement cette distinction qu'on n'a pas prise en considération dans notre hypothèse, vu le manque d'information.

La classification n'est qu'une étape dont le but est de trouver des sous-ensembles homogènes en vue de l'application des méthodes de prédiction.

Nous poursuivons notre démarche en formulant les deux modèles de connaissances de sélection de mots-clés par les fonctions génériques déjà existantes afin d'améliorer leurs paramètres. Des tests effectués ont révélé une amélioration de l'ordre de 10 % dans les résultats de notre méthode par rapport aux résultats de la méthode existante.

Perspective

La catégorisation des mots-clés utilisés par les internautes est un domaine d'investigation récent. La structure même des bases de données n'est pas conçue pour le concept des moteurs de recherches. En effet, certaines variables, telles que le coefficient d'indexation et le coût par clic réel du mot-clé, ne sont pas collectées et peuvent néanmoins jouer un rôle déterminant dans la signification des clusters. Évidemment, le réaménagement de la structure de bases de données sera très coûteux.

L'exploitation de l'approche sémantique reste toujours en étude et exige des techniques d'apprentissage. Cette recherche peut-être améliorée par les outils d'indexation textuelle donnant ainsi plus de sémantique aux groupes des mots-clés. Mais il reste que l'interprétation du mot-clé par différents lecteurs demeure subjective.

BIBLIOGRAPHIE

Abonyi, J., & Feil, B. (2007). *Cluster Analysis for Data Mining and System Identification*. (Édit. Birkhäuser Verlag) Berlin, Allemagne.

Aloise, D., Deshpande, A., Hansen, P., & Popat, P. (2009). *NP-hardness of Euclidean sum-of-squares clustering*. Machine Learning, 75, p. 245-249. doi:10.1007/s10994-009-5103-0

Baeza-Yates, R., Hurtado, C., Mendoza, M., & Dupret, G. (2005). Modeling user search behavior. *Web Congress, 2005*. Mideplan, Chile. doi: 10.1109/LAWEB.2005.23

Baya, A. E., & Granitto, P. M. (2013). How Many Clusters: A Validation Index for Arbitrary-Shaped Clusters. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(2), p. 401-414. doi: 10.1109/tcbb.2013.32

Benzécri, J. P., & Benzécri, F. (1980). *Pratique de l'Analyse des données*. (2e éd.). Paris, France : Dunod.

Bezdek, J. C. (1981). *Pattern Recognition using the Fuzzy c-means technique*. Plenum Press, New York, 1981.

Bouveyron, C. (2012). *Contributions à l'apprentissage statistique en grande dimension, adaptatif et sur données atypiques*. (Mémoire, Université Paris 1 Panthéon-Sorbonne, Paris, France). Tiré de http://samm.univ-paris1.fr/IMG/pdf/HDR_Bouveyron.pdf

Bouveyron, C. (2013). *Data mining et analyse de données : Apprendre et décider à partir de données*. Supports de cours, p. 63, France : Université Paris 1 Panthéon-Sorbonne. Tiré de http://fr.scribd.com/doc/202514903/DM-Cours-Final

Bradley, E., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, USA: Chapman & Hall.

Bradley, P. S., Fayyad, U. M., & Reina, C. A. (1998). Scaling EM Clustering to Large Databases. *Microsoft Research Report*, MSR-TR-98-35. Redmond, Washington: USA, p. 9-15. Tiré de http://parkcu.com/blog/wp-content/uploads/2013/07/Bradley-Fayyad-Reina-1998-Scaling-EM-Expectation-Maximization-Clustering-to-Large-Databases.pdf

Burriel, S. (2010). *Google AdWords. Scénario complet pour réussir sa campagne marketing*. Ed. Pearson, 46, Paris, France.

Chan, N., Adjengue, L., Gamache, M., Marcotte, P., & Savard, G. (2008). *Rapport final: Phase de faisabilité*. Montréal, QC : École Polytechnique de Montréal.

Clément, B. (2013). *Introduction au data mining avec Statistica*. [Présentation Power Point]. Tiré de http://www.groupes.polymtl.ca/mth6301/mth8302/DataMining-partie1-Introduction.pdf

Dans *Wikipédia*. Consulté le 20 janvier 2014, tiré de http://fr.wikipedia.org/wiki/Apprentissage_automatique

Dempster, A.P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1), p. 1-38. Tiré de http://web.mit.edu/6.435/www/Dempster77.pdf

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Unsupervised learning and clustering. Dans *Pattern classification* (2e éd., p. 517-598).

Duijndam, K. (2011). *AdWords bid optimisation*. (Thèse doctorat, VU University Amsterdam, NL). Tiré de https://www.few.vu.nl/en/Images/stageverslag-duijndam_tcm39-234413.pdf

Fayyad, U., Piatetsky-Shiparo, G., & Smyth, P. (1996). *From data mining to knowledge discovery in databases*. (AI magazine). Tiré de http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

Friedman, J. H., & Meulman, J. J. (2004). Clustering objects on subsets of attributes. *Journal of the Royal Statistical Society Series B,* 66(4), p. 815–849. Tiré de http://www.datatheory.nl/pages/Friedman%26Meulman.pdf

Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: Theory, Algorithms, and applications*. Tiré de http://dx.doi.org/10.1137/1.9780898718348

Gath, I., & Geva, A. B. (1989). *Unsupervised optimal fuzzy clustering*. IEEE Trans. Pattern Analysis and Machine Intelligence. Tiré de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.2648&rep=rep1&type=pdf

Google (2014). Au cœur de la recherche. Consulté le 25 janvier 2014, tiré de http://www.google.com/intl/fr/insidesearch/howsearchworks

Hartigan, J.A., & Wong, M.A. (1979). Algorithm AS 136. A k-means clustering algorithm. *Applied Statistics*. 28(1), p. 100-108. Tiré de http://www.jstor.org/stable/2346830

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, Inference and prediction*. (2e éd.). Tiré de http://statweb.stanford.edu/~tibs/ElemStatLearn/

Hosanagar, K., & Abhishek, V. (2012). Optimal bidding in multi-Item multi-Slot sponsored search auctions. *Operations research,* 61(4), p. 855-873. Tiré de http://dx.doi.org/10.2139/ssrn.1544580

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, Inc., New York, 1990.

Kaufman, L., & Rousseeuw, P. J. (1988). Partitioning Around Medoids. Dans *Finding groups in data: An introduction to clusters analysis*. p. 68-125. doi: 10.1002/9780470316801-ch2

Kitts, B., & Leblanc, B. (2004). Optimal bidding on keyword auctions. *Electronic Markets*. 14(3), p. 186-201. doi: 10.1080/1019678042000245119

Kitts, B., Laxminarayan, P., LeBlanc, B., & Meech, R. (2005). A formal analysis of search auctions including predictions on click fraud and bidding tactics. *ACM Conference on Electronic Commerce*. Vancouver, UK. Tiré de http://www.appliedaisystems.com/papers/ECom_Paper17.pdf.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. Édité par Cam, M. L., et Neyman J., B. University of California, Press 1967, V. 1, p. 281-297.

Matusevich, D. S., Ordonez, C., & Baladandayuthapani, V. (2013). A Fast Convergence Clustering Algorithm Merging MCMC And EM Methods. *Proc. of the 22nd ACM international conference*, p. 1525-1528. Tiré de http://www2.cs.uh.edu/~ordonez/w-2013-CIKM-famcem.pdf

Mohebbi, M. (25 mai 2011). *Mining patterns in search data with Google Correlate*. Tiré de http://googleblog.blogspot.ca/2011/05/mining-patterns-in-search-data-with.html

Netcraft (2013). Web serveur survey. Consulté le 20 janvier 2014, tiré de http://news.netcraft.com/archives/2013/11/01/november-2013-web-server-survey.html

Partovi Nia, V., & Davison, A. C. (2012). High-Dimensional Bayesian Clustering with Variable Selection: The R Package bclust. *Journal of Statistical Software*, 47(5), p.1-22. Tiré de http://www.jstatsoft.org/v47/i05/

Quinn, P. (2011). *Modélisation et prédiction du comportement de mots-clés dans des campagnes publicitaires sur les moteurs de recherche*. (Mémoire de maitrise, École polytechnique de Montréal, Montréal, QC).

Réseau Ontario des entreprises, affaires et économie (2014). Consulté le 02 février 2014, tiré de https://www.ontario.ca/fr/affaires-et-economie/la-publicite-efficace-sur-internet

Rusmevichientong, P., & Williamson, D. P. (2006). An adaptive algorithm for selecting profitable keywords for search-based advertising services. *In Proc. 7th ACM conference on Electronic Commerce,* p. 260-269. New York, États-Unis: ACM.

Saad, Y. (2012). Linear algebra methods for data mining with applications to materials. *Conference SIAM 2012 annuel meeting*. University of Minnesota. Tiré de http://www-users.cs.umn.edu/~saad/PDF/SIAM_AN_12.pdf

Saporta, G. (2006). Épidémiologie et data mining ou fouille de données. Dans Académie des sciences (Édit.), *L'épidémiologie humaine : conditions de son développement en France et rôle des mathématiques.* (p. 129-135). Tiré de http://www.bibsciences.org/bibsup/acad-sc/common/articles/rapport3.pdf

Scepi, G. (2010). Clustering algorithms for large temporal data sets. Dans F. Palumbo et al. (Édit.), *Data Analysis and Classification, Studies in Classification, Data Analysis, and Knowledge Organization* (p. 371). doi: 10.1007/978-3-642-03739-9

Statista (2013). The statistics portail. Consulté le 02 février 2014, tiré de http://www.statista.com/topics/1176/online-advertising/chart/1409/global-online-ad-revenue

Suneetha, M., Fatima, S. S., & Pervez, S. M. Z. (2011). Clustering of Web Search Results using Suffix Tree Algorithm and Avoidance of Repetition of same Images in Search Results using L-Point Comparison Algorithm. *Emerging Trends in Electrical and Computer Technology (ICETECT), 2011 International Conference,* p. 1041-1046. Tamil Nadu: Inde. doi : 10.1109/ICETECT.2011.5760272

Tibshirani, R., Walther, G., & Hastie, T. (2000). Estimating the number of clusters in a data set via the gap statistic. *Journal of the royal statistical society: series B*, 63(2), p. 411-423. doi: 10.1111/1467-9868.00293

Tufféry, S. (2007). *Data mining et statistique décisionnelle : l'intelligence des données*. (Nouv. éd. rev. et augm.). Paris, France : Technip.

Wang, X., Qiu, W., & Zamar, R. H. (2007). CLUES : A non-parametric clustering method based on local shrinking. *Computational Statistics & Data Analysis*, 52(1), p. 286-298. doi:10.1016/j.csda.2006.12.016

Witten, D. M., & Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490), p. 713–726. doi:10.1198/jasa.2010.tm09415

Wu, X., & Kumar, V. (2009). *The top ten algorithms in data mining*. Séries: Chapman & Hall / CRC Data mining and knowledge discovery series. (9), p. 21-36.

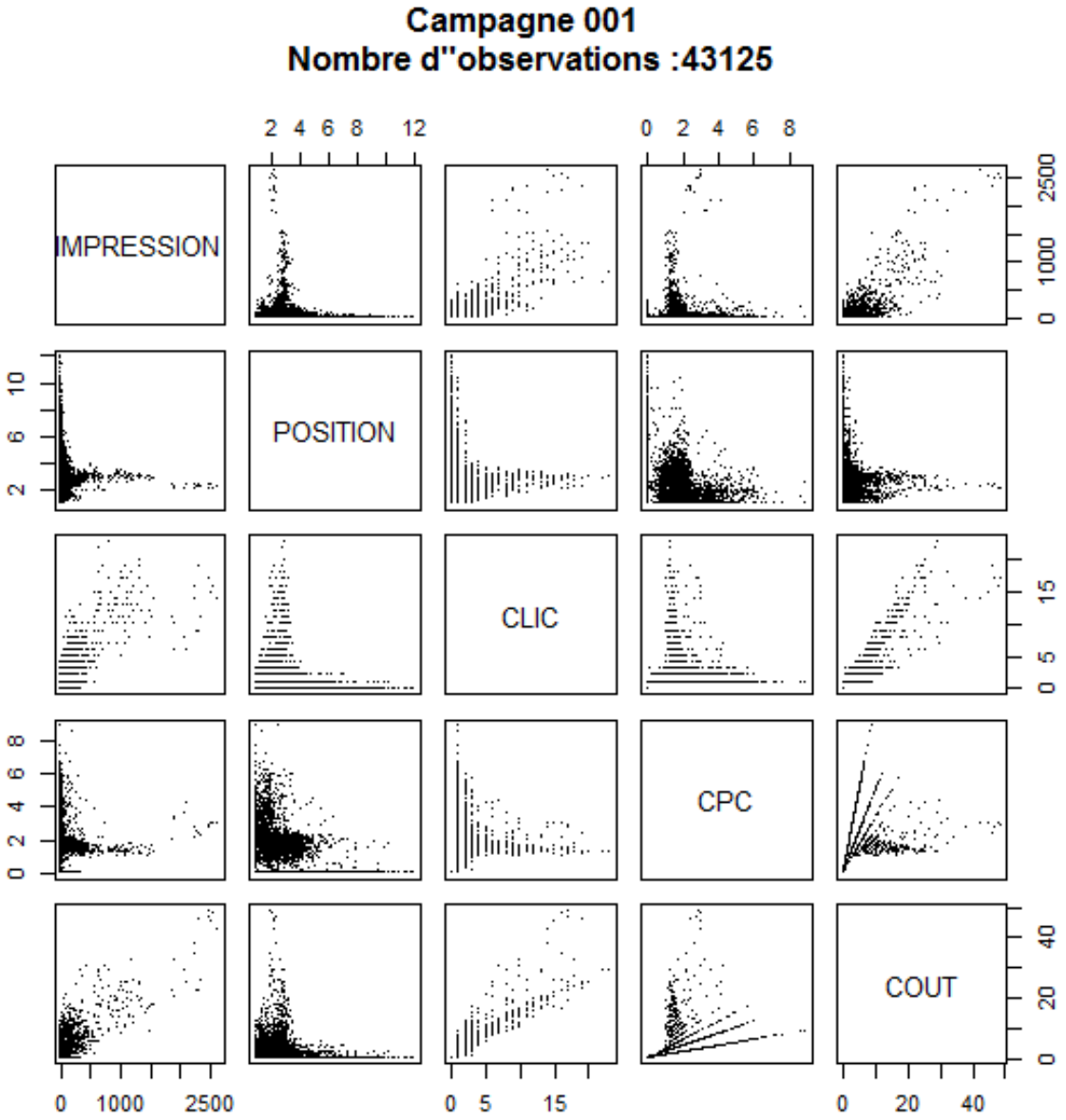
Zaki, M. J., & MEIRA, W. (2013). *Data Mining and Analysis: Fundamental concept and Algorithms*. p. 370. Tiré de http://www.cs.rpi.edu/~zaki/PaperDir/DMABOOK.pdf

Zhang, J., & Dimitroff, A. (2003). *The impact of webpage content characteristics on webpage visibility in search engine results*. Consulté le 15 janvier 2014, tiré de https://pantherfile.uwm.edu/jzhang/www/publications_files/P2005_3.pdf

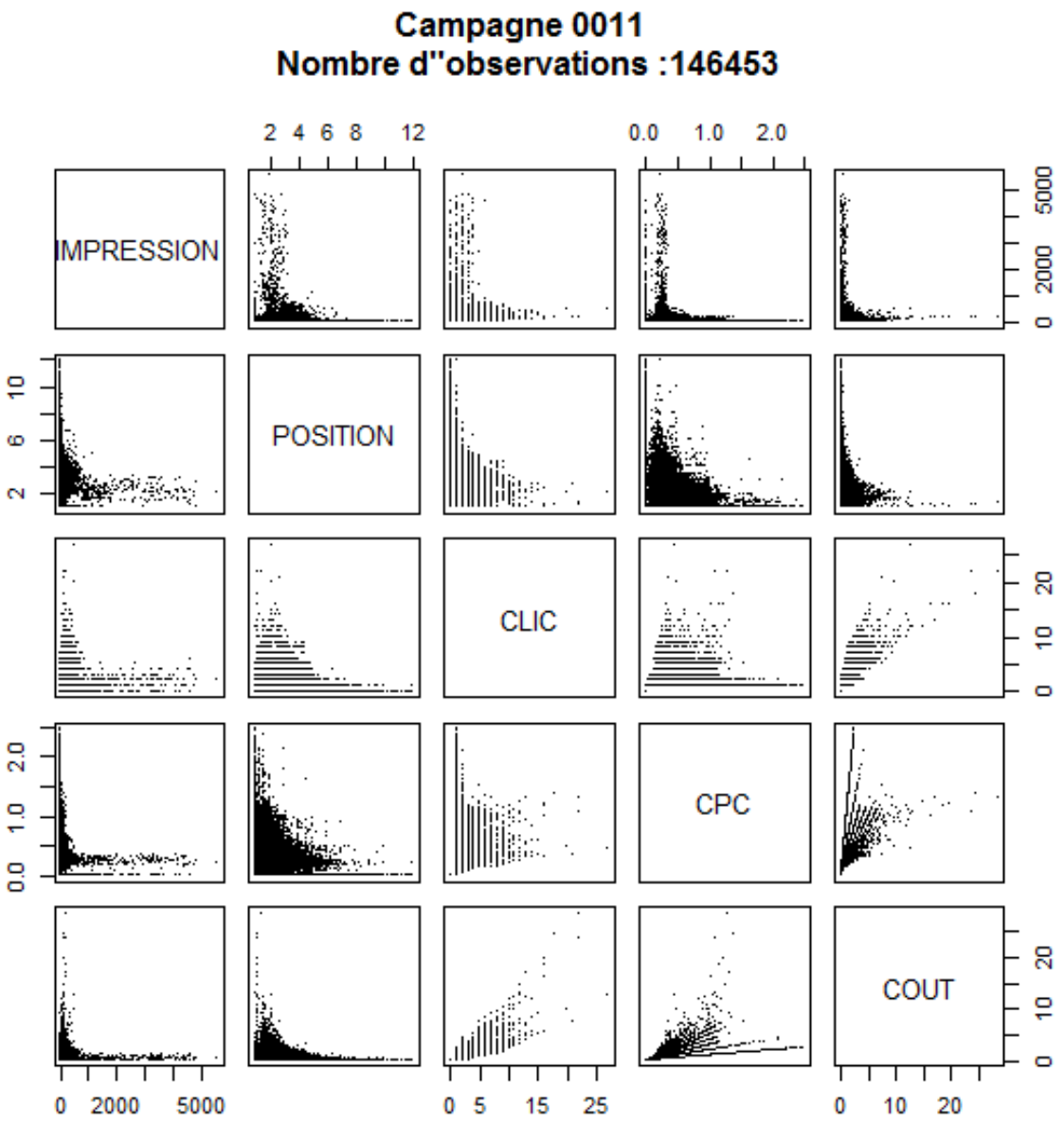
ANNEXE A : EXTRAIT D'ÉCHANTILLON DE CAMPAGNES PUBLICITAIRES

				Données totales (clic = 0 et clic > 0)							Données seulement avec clic > 0					
N°	Agence	Campagne	Durée de campagne	Nombre de mots-clés	Nombre d'observations	Impression Total	Position Pondérée	Clic Total	Coût Total ($)	Conversion total	Nombre de mots-clés	Nombre observations	Impression total	Position Pondérée	Nbre Obs. /Conversion	Taux conversion/ Observations
1	2482	327164	350	309	43125	442161	2,76	6824	12922,56	0	208	3998	240095	2,66	0	0,00%
2	3219	493774	642	228	93634	1908907	1,87	191515	31670,32	127	209	30062	1571413	1,78	98	0,33%
3	2444	406304	622	616	51890	116703	1,86	3954	12079,36	18	397	3725	11868	1,63	18	0,48%
4	10910	576024	721	385	49910	1186923	6,2	3541	25799,57	0	184	2763	373498	5,33	0	0,00%
5	4169	478994	3116	865	275357	7670052	4,23	173188	123480,76	34	589	58080	5701927	4,2	2	0,00%
6	4339	512394	228	1078	52311	371841	2,36	8191	1962,81	5	498	6304	117314	2,33	5	0,08%
7	4249	402874	399	472	63542	566458	2,9	7676	20168,68	0	254	6165	184104	2,64	0	0,00%
8	3219	150944	1565	408	51786	1179041	2,53	54267	12659,39	0	348	12416	867113	2,43	0	0,00%
9	3609	502864	523	320	99592	6162340	4,23	85219	137084,38	486	238	21724	5184351	3,98	406	1,87%
10	6000	568384	1720	144	55613	1052839	2,73	19008	155508,82	62	106	10085	607679	2,39	62	0,61%
11	4299	230924	1707	255	146453	2017381	2,58	30427	13908,57	370	201	18692	1166248	2,47	360	1,93%
12	2529	465424	413	344	56271	1232881	2	60078	54578,67	0	234	16340	934927	1,83	0	0,00%
13	4169	501444	248	5012	306608	3303178	4,87	47212	40126,93	247	1867	21396	1518052	4,15	169	0,79%
14	6570	473774	1180	137	61960	774118	3,36	7248	3185,12	23	109	6016	305464	3,13	22	0,37%
15	2518	81624	859	252	88086	5376271	3,15	39018	26065,95	0	182	12885	4597479	3,08	0	0,00%
16	2518	114904	378	740	78897	372327	4,92	5944	4346,67	123	355	3788	106023	4,29	73	1,93%
17	2518	81304	1606	672	231256	6302948	2,73	289780	334491,86	0	597	76216	5208184	2,53	0	0,00%
18	2518	121564	593	2788	347870	2631502	2,43	58976	99052,7	0	1215	33874	1270951	2,3	0	0,00%
19	10910	576814	343	1148	64004	4027924	2,99	1066	2041,47	22	136	733	2063050	2,99	20	2,73%
20	2444	60334	2090	1102	117248	813651	3,26	8372	16282,6	54	573	7554	128470	3,15	54	0,71%
21	4249	302874	1447	468	181908	1548933	2,64	39467	57940,37	432	282	20378	638547	2,38	334	1,64%
22	2518	108914	897	333	55390	264596	4,63	7099	3053,76	0	208	5424	78060	3,86	0	0,00%
23	4299	230834	1632	315	242165	2377977	3,04	30300	13570,12	229	251	18426	1142192	2,83	225	1,22%
24	2482	158774	389	1509	65851	290217	2,48	4453	6082,97	0	523	2222	99127	1,79	0	0,00%
25	2529	77024	1216	485	159640	7856445	1,59	474931	253115,59	0	315	63172	7223969	1,57	0	0,00%
26	3609	428844	238	599	54555	515305	6,26	573	4300,11	7	105	527	35695	5,47	7	1,33%
27	2482	84784	881	627	51492	139342	2,99	3414	10064,3	28	330	2968	20732	2,58	28	0,94%
28	4169	277694	3215	35	70582	1889022	2,13	87497	35288,85	938	33	29912	1485868	1,92	637	2,13%
29	2518	121134	486	2258	202672	653978	2,08	8749	15015,58	0	488	7909	56107	1,66	0	0,00%
30	6570	474264	1572	131	204570	3015859	3,02	64184	43920,58	321	123	36297	1678239	2,69	317	0,87%

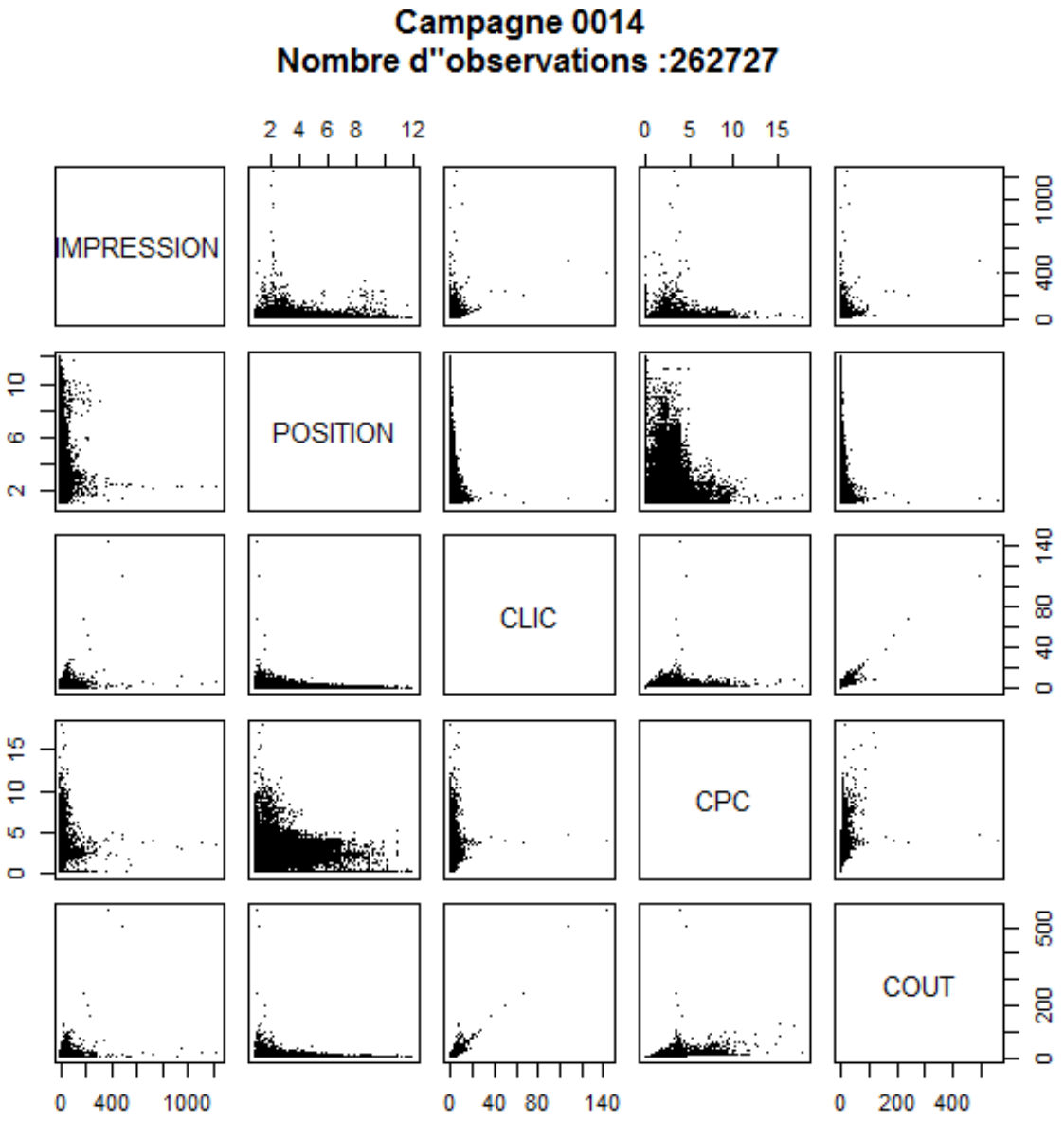
ANNEXE B : EXEMPLES GRAPHIQUES ILLUSTRANT LES RELATIONS ENTRE LES VARIABLES DES CAMPAGNES PUBLICITAIRES

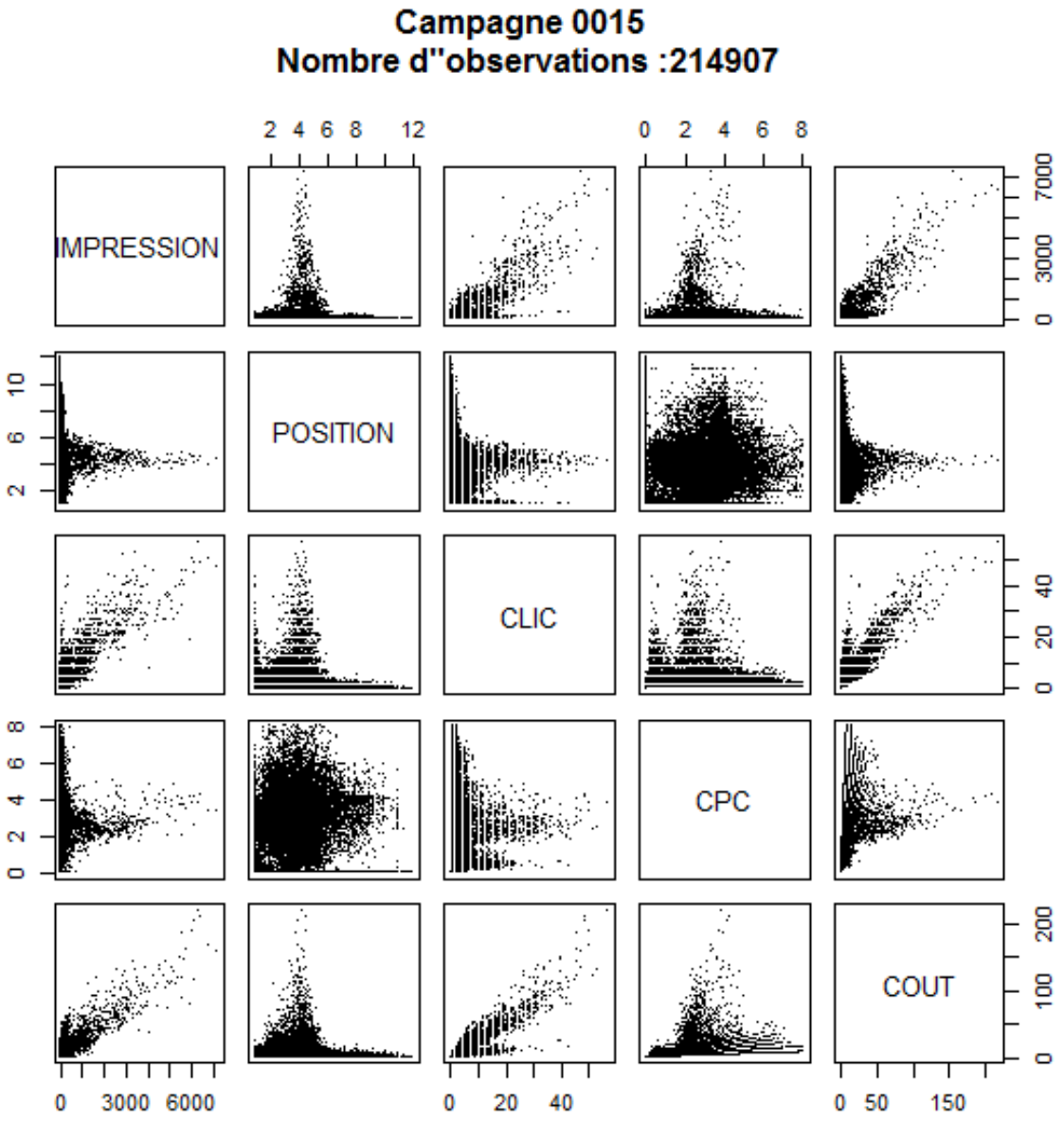


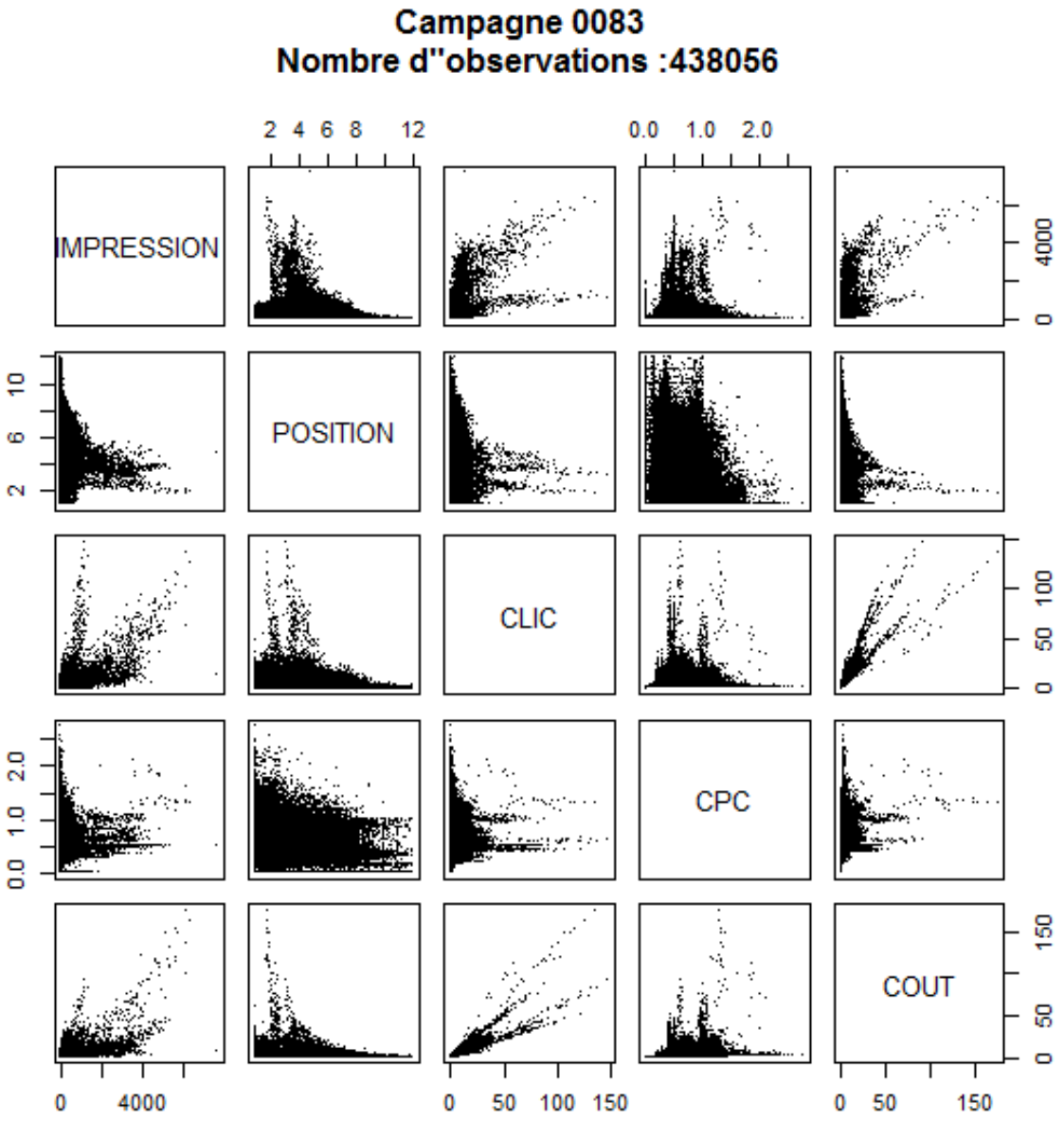
On remarque que cette campagne publicitaire présente peut de clic



Un cas de campagne publicitaire ayant peut de clic.


Campagne 0014
Nombre d''observations :262727
IMPRESSION
POSITION
CLIC
CPC
COUT
0 400 1000
2 6 10
0 40 80 140
0 5 10 15
0 200 400 500
2 4 6 8 12
0 5 10 15
0 400 1000
0 40 80 140
0 200 400


Campagne 0015
Nombre d''observations :214907
IMPRESSION
POSITION
CLIC
CPC
COUT
2 4 6 8 12
0 2 4 6 8
0 3000 7000
2 6 10
0 20 40
0 2 4 6 8
0 50 200
0 3000 6000
0 20 40
0 50 150


Campagne 0083
Nombre d''observations :438056
IMPRESSION
POSITION
CLIC
CPC
COUT
2 4 6 8 12
0.0 1.0 2.0
0 4000
2 6 10
0 50 100
0.0 1.0 2.0
0 50 150
0 4000
0 50 100 150
0 50 150