| | |
|---|---|
| **Titre:** Title: | Modélisation et prédiction du comportement de mots-clés dans des campagnes publicitaires sur les moteurs de recherche |
| **Auteur:** Author: | Patrick Quinn |
| **Date:** | 2011 |
| **Type:** | Mémoire ou thèse / Dissertation or Thesis |
| **Référence:** Citation: | Quinn, P. (2011). Modélisation et prédiction du comportement de mots-clés dans des campagnes publicitaires sur les moteurs de recherche [Mémoire de maîtrise, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/565/ |



| | |
|---|---|
| **URL de PolyPublie:** PolyPublie URL: | https://publications.polymtl.ca/565/ |
| **Directeurs de recherche:** Advisors: | Michel Gamache, & Gilles Savard |
| **Programme:** Program: | Génie industriel |

UNIVERSITÉ DE MONTRÉAL

# MODÉLISATION ET PRÉDICTION DU COMPORTEMENT DE MOTS-CLÉS DANS DES CAMPAGNES PUBLICITAIRES SUR LES MOTEURS DE RECHERCHE

PATRICK QUINN

DÉPARTEMENT DE MATHÉMATIQUES ET GÉNIE INDUSTRIEL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION

DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES

(MATHÉMATIQUES APPLIQUÉES)

AVRIL 2011

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé:

**MODÉLISATION ET PRÉDICTION DU COMPORTEMENT DE MOTS-CLÉS DANS DES CAMPAGNES PUBLICITAIRES SUR LES MOTEURS DE RECHERCHE**

Présenté par : QUINN Patrick

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

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

M. ADJENGUE Luc, Ph. D., président

M. HERTZ Alain, D. Sc., membre

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

M. SAVARD Gilles, Ph. D., membre et codirecteur de recherche

*If we knew what we were doing it wouldn't be research.*

-Albert Einstein

# REMERCIEMENTS

D'abord, je tiens à remercier mon directeur de recherche, Michel Gamache. Son expertise, sa disponibilité et son dévouement exceptionnel ont permis de structurer mon travail tout au long du projet. Malgré son horaire chargé, il a toujours pris le temps de répondre à mes questions, me donner des conseils et me guider dans ma démarche. Grâce à lui, une relation de collaboration entre l'École Polytechnique et Acquisio s'est développée, permettant à moi-même ainsi que plusieurs autres étudiants d'appliquer des notions de statistique et d'optimisation dans un milieu de travail concret et stimulant.

J'aimerais aussi remercier Sandrine Paroz, ma collègue de travail chez Acquisio. Nous avons été initiés au monde du marketing sur les moteurs de recherche en même temps, mais Sandrine a su me guider par son expertise en optimisation ainsi que par son approche scientifique rigoureuse et méthodique. Elle m'a fourni un excellent encadrement pendant chacune des phases du projet et n'a jamais hésité à investir son temps dans le but de faciliter mon cheminement.

Lorsque le projet semblait stagner et les pistes de recherche devenaient un peu moins évidentes, Gilles Savard, mon codirecteur de recherche, a su nous guider dans la bonne direction. Ces conseils fournis aux moments opportuns étaient très appréciés.

Je suis également très reconnaissant envers Acquisio, qui m'a fourni un milieu de travail agréable, enrichissant et motivant. Je remercie tous les gens d'Acquisio de m'avoir accueilli dans leurs bureaux et de m'avoir fourni tout le matériel et l'encadrement nécessaire à l'accomplissement de mon projet. Plus particulièrement, je tiens à remercier Richard Couture, qui a eu la vision et l'initiative nécessaire pour démarrer le projet d'optimisation sur lequel j'ai travaillé. Lui, ainsi que Martin Mailloux, Marc Poirier et Martin Le Sauteur m'ont fourni un excellent encadrement en participant régulièrement aux rencontres de suivi du projet. Je dois également remercier tous les autres employés d'Acquisio qui m'ont constamment fourni de l'aide, autant pour des questions de support technique que pour l'extraction des données, ou simplement pour des questions relatives au fonctionnement du domaine. Sans leur collaboration, mon expérience de recherche aurait été beaucoup plus difficile.

Finalement, j'aimerais remercier Acquisio, le FQRNT et le CRSNG pour leur contribution financière à la bourse BMP-Innovation qui m'a été octroyée. Ce support financier généreux m'a permis de me consacrer entièrement à mon travail de recherche pendant les deux dernières années.

# CONFIDENTIALITÉ

La recherche présentée dans ce mémoire a été effectuée chez Acquisio, une entreprise qui développe et commercialise une plateforme logicielle visant à gérer des campagnes publicitaires sur les moteurs de recherche, sur les réseaux sociaux et avec les mécanismes de type « Ad Exchange ». Les méthodes et algorithmes ont été développés en vue d'être intégrés à la plateforme logicielle, afin de permettre aux clients de mieux gérer leurs campagnes publicitaires. Puisque le domaine est très compétitif, Acquisio souhaite que certaines informations demeurent confidentielles.

Acquisio déposera, sous peu, un brevet visant à protéger toute la propriété intellectuelle relative à sa plateforme. Certains des algorithmes présentés dans ce mémoire seront protégés par ce brevet et peuvent donc être expliqués sans omettre aucune information. Cependant, dans les cas où l'information n'a pas pu être entièrement protégée, nous excluons intentionnellement quelques détails. Par exemple, nous présentons parfois des algorithmes en ne précisant pas les valeurs exactes de paramètres qui ont été fixées. Afin de faciliter la compréhension du lecteur, nous mentionnons toutefois des exemples de valeurs qui pourraient être utilisées.

Acquisio souhaite également que les données utilisées pour effectuer nos analyses demeurent confidentielles. Nous ne mentionnons donc pas les clients qui sont associés aux banques de données utilisées. De plus, nous ne pouvons pas fournir tous les détails sur les mots-clés qui constituent les ensembles étudiés. Cependant, nous jugeons que cela ne devrait pas affecter la clarté des explications fournies.

# RÉSUMÉ

Les moteurs de recherche utilisent des mécanismes d'enchères très sophistiqués pour déterminer le positionnement des annonces textuelles sur leurs pages de résultats. Tous les annonceurs désirant faire la promotion de leurs produits ou services doivent se faire compétition au sein d'une enchère fermée visant à établir l'ordre dans lequel leurs annonces sont présentées. Plus précisément, l'ordre d'apparition des annonces est déterminé en fonction de la valeur que les annonceurs sont prêts à payer par clic, ainsi que la pertinence de l'annonce et du site Web qui lui est associé. Généralement, les positions au haut de la liste obtiennent plus de visibilité et plus de clics que les positions au bas de la liste. Les annonceurs cherchent donc à fixer leurs valeurs d'enchère de façon à obtenir une visibilité et un nombre de clics satisfaisants, en limitant toutefois leurs coûts à des montants raisonnables. En appliquant ce principe à plusieurs milliers de mots-clés tout en considérant les diverses contraintes relatives au problème, il peut devenir extrêmement difficile de gérer une campagne publicitaire de façon efficace.

Notre travail vise essentiellement à développer des algorithmes automatisables permettant d'améliorer le rendement des campagnes d'annonces textuelles sur les moteurs de recherche. Pour y arriver, nous présentons d'abord une synthèse qui explique le fonctionnement de ce milieu encore très peu connu. Par la suite, en s'inspirant des travaux publiés dans la littérature, nous modélisons le problème à l'aide d'un programme linéaire. En supposant que la performance d'une annonce peut être mesurée par le nombre de clics qu'elle obtient, le modèle vise à déterminer une façon optimale d'affecter les annonces aux diverses positions disponibles, tout en respectant une contrainte de budget. L'utilisation de fonctions de prédiction qui estiment le nombre de clics et le coût par clic moyen en fonction de la position moyenne de l'annonce est nécessaire pour chacun des mots-clés considérés.

Puisque les mots-clés ne possèdent pas tous le même potentiel de prédiction, nous développons un algorithme de classification qui vise à affecter chacun de ces mots-clés à une méthode ou un traitement spécifique. En fonction de leurs caractéristiques, les mots-clés sont classés de façon à réduire les erreurs de prédiction et ainsi maximiser la portée du modèle d'optimisation.

Lorsque les données historiques le permettent, nous utilisons des méthodes de régression pour prédire le comportement des mots-clés en fonction de la position moyenne. Cependant, nos essais démontrent qu'une grande proportion des mots-clés ne fournissent pas des régressions

statistiquement acceptables, ce qui met en évidence le besoin de développer des méthodes de prédiction alternatives. Nous étudions alors la possibilité d'utiliser des fonctions de prédiction génériques, soit des fonctions qui supposent que les taux de décroissance relatifs du nombre de clics et du coût par clic moyen sont presque constants d'un mot-clé à l'autre. Suite à des analyses effectuées à partir des banques de données à notre disposition, nous concluons que cette approche génère des estimations suffisamment précises pour être utilisée à des fins de prédiction. En comparant l'approche de décroissance linéaire avec celle de décroissance exponentielle, nous choisissons de retenir la méthode exponentielle.

Nous procédons alors à plusieurs essais dans le but de raffiner la méthode et ainsi améliorer la qualité des prédictions. Nous constatons que le repositionnement dynamique des fonctions de prédiction et la pondération décroissante des données en fonction de leur ancienneté permettent de diminuer considérablement les marges d'erreur associées aux prédictions. Une fois ces améliorations ajoutées à la méthode, les erreurs de prédiction atteignent des seuils que nous considérons satisfaisants. Toutefois, nous proposons plusieurs pistes de recherche qui pourraient possiblement améliorer davantage les résultats obtenus.

Globalement, ce projet de recherche nous a permis d'acquérir une meilleure compréhension des mécanismes utilisés dans le milieu du marketing sur les moteurs de recherche. Nos nombreuses analyses ont permis d'approfondir plusieurs concepts et vérifier certaines hypothèses. Notre modèle d'optimisation, notre algorithme de classification ainsi que les méthodes de prédiction que nous proposons pourront éventuellement être intégrés à une plateforme logicielle de façon à constituer un algorithme d'optimisation complet et fonctionnel, utilisable sur une base quotidienne. Un tel algorithme permettrait aux annonceurs de gérer leurs campagnes publicitaires de façon beaucoup plus efficace et améliorerait probablement leur profitabilité.

# ABSTRACT

Search engines use sophisticated auction mechanisms to determine the positioning of text ads on their result pages. Advertisers seeking to promote their products or services must compete in sealed auctions to establish the order in which their ads will appear. More specifically, ad ranks are calculated based on the amounts the advertisers are willing to pay for each click, as well as the relevance of the ads and their related websites. Generally, ads placed at the top of the results list get more visibility and generate more clicks than the ads at the bottom of the list. Therefore, advertisers try to find the optimal bids that will allow them to obtain sufficient visibility and clicks, while maintaining reasonable costs.

Our main objective is to develop automatable algorithms that can increase the performance of search engine text ad campaigns. In order to achieve this, we first present a summary explaining the various aspects of this relatively new research field. Thereafter, based on several studies that have been published, we model the problem using a linear programming approach. Assuming an ad's performance is measured by the number of clicks it generates, our linear program is designed to determine an optimal way to allocate the text ads to the various available positions while respecting a budget constraint. In order to estimate the number of clicks and the average cost per click for each position, the use of prediction functions is necessary.

Since keywords do not all provide the same quality of regressions, we develop a classification algorithm that allows us to identify which processing can be used, depending on each individual keyword's characteristics. We aim to classify the keywords in a way that should reduce, as much as possible, the prediction errors. This should also maximize the scope of the optimization model, allowing us to estimate the clicks and costs per click of almost every single keyword in a campaign.

When we are able to obtain adequate regressions using the keywords' historical data, it is relatively easy to predict clicks and cost per click as a function of position. However, our analysis shows that a large proportion of keywords do not provide statistically acceptable regressions, which makes it necessary to develop alternative prediction methods. We then study the possibility of using generic prediction functions. Such a prediction approach suggests that the relative decrease rate of the click and cost per click functions are almost constant from one keyword to another. After several tests performed on our databases, we conclude that the generic

functions provide estimations that are precise enough to be used for prediction purposes. While comparing a linear function approach with an exponential function approach, we determine that the exponential version offers more advantages.

We then seek to refine our method, in order to ultimately improve the quality of predictions we are able to generate. We find that the dynamic repositioning of our prediction functions and the use of an exponentially decreasing weight in order to prioritize recent observations allow us to achieve much better results. Once these improvements are added to the method, we are quite satisfied with the prediction errors that are obtained. However, we still suggest several research areas that could potentially further improve the quality of the predictions.

Overall, this research project has allowed us to acquire a better understanding of the mechanisms that are used by search engines to manage their publicity networks. The optimization model, the classification algorithm and the prediction methods we suggest should eventually be integrated in a software platform, in such a way as to form a complete and functional optimization algorithm that can be used on a daily basis. Such an algorithm would allow advertisers to manage their text ad campaigns more efficiently and should increase their profitability.

# TABLE DES MATIÈRES

REMERCIEMENTS ..... IV

CONFIDENTIALITÉ ..... VI

RÉSUMÉ ..... VII

ABSTRACT ..... IX

TABLE DES MATIÈRES ..... XI

LISTE DES TABLEAUX ..... XIV

LISTE DES FIGURES ..... XVI

LISTE DES SIGLES ET ABRÉVIATIONS ..... XVIII

LISTE DES ANNEXES ..... XIX

CHAPITRE 1 INTRODUCTION ..... 1

1.1 Les différents types de publicité ..... 2

1.2 Définitions importantes ..... 3

1.3 Historique des méthodes de classement des annonces ..... 15

1.4 L'effet de la position ..... 17

1.4.1 Les études publiées ..... 17

1.4.2 Nos analyses ..... 18

1.4.3 Les raisons qui expliquent les taux de clic décroissants ..... 20

1.4.4 Le raisonnement des moteurs de recherche ..... 21

1.4.5 La profitabilité de chaque position ..... 21

1.4.6 Conclusion ..... 23

1.5 L'algorithme de classement des annonces ..... 23

1.6 La gestion d'une campagne ..... 25

1.6.1 Les objectifs ..... 25

1.6.2 La gestion des enchères ..... 26
1.7 L'importance des mathématiques ..... 26
1.8 Les intérêts de chacun des joueurs ..... 27
1.9 Présentation de l'entreprise ..... 29
1.9.1 Les agences publicitaires ..... 29
1.9.2 Les logiciels de gestion de campagnes ..... 30
1.9.3 Acquisio ..... 30
1.10 Présentation du sujet de recherche ..... 31
CHAPITRE 2 REVUE DE LITTÉRATURE ..... 33
2.1 Quelques modélisations intéressantes ..... 34
2.2 Méthodes à valeur d'enchère et position constantes ..... 35
2.3 Méthodes à CPC et positions multiples ..... 37
CHAPITRE 3 MODÈLE D'OPTIMISATION ..... 40
3.1 Structure des campagnes publicitaires ..... 40
3.2 Modèle initial ..... 42
3.3 Limites du modèle ..... 48
3.4 Modèle alternatif ..... 51
CHAPITRE 4 CLASSIFICATION DES MOTS-CLÉS ..... 61
4.1 Problèmes rencontrés ..... 61
4.2 Algorithme de classification ..... 68
CHAPITRE 5 FONCTIONS DE PRÉDICTION GÉNÉRIQUES ..... 80
5.1 Principe ..... 80
5.2 Données utilisées pour les analyses ..... 82
5.3 Fonctions génériques linéaires ..... 85

5.4 Fonctions génériques exponentielles ..... 99
5.5 Comparaison des deux approches ..... 109
5.6 Raffinement de la méthode exponentielle ..... 114
5.7 Conclusion ..... 121
CONCLUSION ..... 124
ANNEXES ..... 135

# LISTE DES TABLEAUX

Tableau 1.1 : Estimations de rendement par position pour un mot-clé quelconque .....22

Tableau 1.2 : Valeurs d'enchère et indices de qualité.....24

Tableau 1.3 : Classement des annonceurs.....24

Tableau 3.1 : Exemple de discrétisation de fonctions de prédiction (revenus, coûts et valeurs d'enchère en fonction de la position) .....46

Tableau 3.2 : Exemple de données historiques (position moyenne, nombre de clics et nombre de conversions) d'un mot-clé typique.....50

Tableau 3.3 : Exemple de discrétisation de fonctions de prédiction (nombre de clics, CPC moyens et valeurs d'enchère en fonction de la position).....56

Tableau 4.1 : Conditions nécessaires pour prédire le comportement d'un mot-clé à l'aide de régressions.....69

Tableau 5.1 : Description globale des banques de données utilisées pour les analyses.....83

Tableau 5.2 : Répartition des clics et des coûts en fonction des mots-clés pour toutes les banques de données agrégées.....84

Tableau 5.3 : Données historiques de position moyenne, nombre de clics et CPC moyen d'un mot-clé quelconque .....95

Tableau 5.4 : Résultats globaux agrégés sur 20 banques de données pour les critères d'évaluation des fonctions de clics et de CPC moyen .....111

Tableau 5.5 : Données d'un mot-clé qui a subi une forte diminution de volume pendant sa période de prédiction.....115

Tableau 5.6 : Prédictions de clics en fonction de la position moyenne pour les journées 47 à 60 .....116

Tableau 5.7 : Comparaison des résultats de la méthode exponentielle sans repositionnement dynamique (A) et avec repositionnement dynamique (B) .....117

Tableau 5.8 : Poids calculés en fonction des valeurs de $S$ utilisées, pour une période de 30 jours .....119

Tableau 5.9 : Valeurs des critères d'erreur suite à l'application de divers poids exponentiels aux observations.....120

Tableau 5.10 : Comparaison des résultats de la méthode exponentielle sans repositionnement dynamique (A), avec repositionnement dynamique (B) et avec repositionnement dynamique ainsi que pondération des observations (C).....120

Tableau 5.11 : Valeurs des critères d'erreur obtenues pour le mot-clé de la Figure 5.10 en utilisant les régressions comme fonctions de prédiction.....123

# LISTE DES FIGURES

Figure 1.1 : Exemple d'annonce textuelle....9

Figure 1.2 : Exemple des positions attribuées aux annonces par Google ....11

Figure 1.3 : Exemple des positions attribuées aux annonces par Google (cas avec positions Premium)....11

Figure 1.4 : Taux de clic moyens par position (agrégés sur toutes les banques de données) ....20

Figure 3.1 : Organisation des campagnes d'une agence publicitaire ....41

Figure 3.2 : Exemples de fonctions de prédiction : revenus (A), coûts (B) et valeurs d'enchère (C) en fonction de la position moyenne....45

Figure 3.3 : Graphique du nombre de conversions en fonction de la position moyenne pour un mot-clé....51

Figure 3.4 : Graphique du nombre de clics en fonction de la position moyenne pour un mot-clé 53

Figure 3.5 : Exemples de fonctions de prédiction : nombre de clics (A), CPC moyens (B) et valeurs d'enchère (C) en fonction de la position moyenne....55

Figure 4.1 : Exemples de graphiques de nombre de clics (A) et de CPC moyen (B) avec forte dispersion dans les données....62

Figure 4.2 : Exemple de graphique de clics en fonction de la position moyenne pour un mot-clé à faible volume....65

Figure 4.3 : Exemples de graphiques de nombre de clics (A) et de CPC moyen (B) en fonction de la position moyenne, avec plage de données de position trop restreinte....65

Figure 4.4 : Exemples de graphiques de nombre de clics (A) et de CPC moyen (B) en fonction de la position moyenne, avec plage de données de position suffisamment grande ....66

Figure 4.5 : Algorithme de classification ....70

Figure 5.1 : Exemples de graphiques de clics pour deux mots-clés avec taux de décroissance relatifs semblables, avant (A et B) et après (C et D) mise à l'échelle....86

Figure 5.2 : Répartition des pentes mises à l'échelle pour les graphiques de clics en fonction de la position moyenne .....89

Figure 5.3 : Répartition des pentes mises à l'échelle pour les graphiques de CPC moyen en fonction de la position moyenne .....90

Figure 5.4 : Représentation graphique des données historiques de clics (A) et de CPC moyen (B) en fonction de la position moyenne pour un mot-clé quelconque.....95

Figure 5.5 : Représentation graphique des fonctions génériques linéaires pour la prédiction des clics (A) et du CPC moyen (B) en fonction de la position moyenne pour un mot-clé quelconque .....97

Figure 5.6 : Exemple de graphiques de clics pour deux mots-clés (A et B) avec taux de décroissance relatifs semblables.....101

Figure 5.7 : Répartition des coefficients de décroissance $cclics$ pour les graphiques de nombre de clics en fonction de la position moyenne .....103

Figure 5.8 : Répartition des coefficients de décroissance $ccpc$ pour les graphiques de CPC moyen en fonction de la position moyenne.....104

Figure 5.9 : Représentation graphique des fonctions génériques linéaires pour la prédiction des clics (A) et du CPC moyen (B) en fonction de la position moyenne pour un mot-clé quelconque .....107

Figure 5.10 : Exemple de mot-clé pour lequel la régression fournit une fonction de prédiction satisfaisante .....122

# LISTE DES SIGLES ET ABRÉVIATIONS

| | |
|---|---|
| CPC | Coût par clic |
| GSP | « Generalized Second-Price » : expression anglaise utilisée pour désigner l'algorithme de classement et de tarification des annonces utilisé par les principaux moteurs de recherche |
| PPC | « Pay-per-click » : expression anglaise utilisée pour désigner le type d'annonce dans lequel les tarifs publicitaires sont déterminés en fonction du nombre de clics obtenus |
| RSI | Retour sur investissement |
| URL | « Uniform Resource Locator » : expression anglaise utilisée pour désigner une adresse Web |

# LISTE DES ANNEXES

ANNEXE 1 ..... 135

ANNEXE 2 ..... 138

ANNEXE 3 ..... 147

ANNEXE 4 ..... 154

ANNEXE 5 ..... 157

# CHAPITRE 1 INTRODUCTION

Selon le *International Telecommunication Union,* le nombre total d'utilisateurs d'Internet dans le monde n'a cessé de croître au cours des 10 dernières années. En effet, évalué à environ 400 millions de personnes en 2000, ce nombre a connu une croissance relativement constante d'année en année, pour atteindre un niveau de 2,1 milliards en 2010 (International Telecommunication Union [ITU], 2010). Puisque la majorité des gens qui naviguent sur le Web le font à l'aide d'un moteur de recherche, il est facile d'en déduire que l'utilisation des divers moteurs de recherche a également connu une croissance importante au cours des dernières années.

Les moteurs de recherche fournissent un environnement permettant d'effectuer des requêtes et d'obtenir une multitude de résultats pertinents, souvent en une fraction de seconde. Ils permettent aux utilisateurs d'accéder, gratuitement et à n'importe quel moment de la journée, à des milliards de pages Web, images et autres types de fichiers. Ils représentent, sans aucun doute, la source d'information la plus puissante et la plus abondante à notre disposition. Puisque les principaux moteurs de recherche sont toujours accessibles sans frais, ils doivent développer des stratégies leur permettant de générer des revenus. C'est pourquoi ils ont tous recours à des mécanismes de publicité.

Grâce à des systèmes de marketing publicitaire très sophistiqués, les moteurs de recherche réussissent à générer des milliards de dollars en revenus annuellement. Des publicités placées dans des sites Web quelconques sous forme d'images ou de vidéos, ainsi que des annonces textuelles sur les pages de recherche constituent des mécanismes publicitaires très lucratifs, étant donné le nombre d'utilisateurs incroyablement élevé qui naviguent sur le Web à chaque jour.

En 2009, la somme des revenus générés par les moteurs de recherche s'élevait à plus de 37 milliards de dollars. Les quatre plus gros moteurs de recherche en termes de volume, Google, Yahoo, Baidu et Bing affichaient des revenus annuels de 23,7, 6,5, 4,4 et 3,1 milliards de dollars US respectivement (Google Finance, 2010a; Google Finance, 2010b; Google Finance, 2010c; Microsoft Corporation, 2009). La majorité de ces revenus provenaient des divers mécanismes de publicité qui y sont présents. Par exemple, Google affirmait que 99% de ses revenus en 2007 et 97% de ses revenus en 2008 provenaient de son système de publicité (Google Inc., 2008).

Selon une étude du groupe Forrester Research, 6% des ventes au détail aux États-Unis en 2009 ont été effectuées sur le Web et 42% ont été influencées par des recherches sur le Web (Schonfeld, E., 2010). L'étude prédit également que ces valeurs augmenteront de façon relativement constante dans les prochaines années. Devant ces statistiques, il n'est pas difficile de constater que l'industrie des publicités sur les moteurs de recherche occupe une place importante dans notre économie.

## 1.1 Les différents types de publicité

Il existe deux principaux types de publicité associés aux moteurs de recherche : les publicités sur le réseau « Display » et celles sur le réseau « Search ». Il est important de comprendre la distinction entre chacun de ces réseaux, car ce sont des milieux publicitaires qui possèdent des conditions et des contraintes très différentes. Par conséquent, les stratégies de gestion et d'optimisation associées à chacun ne sont pas les mêmes.

Les publicités sur le réseau Display

D'abord, le réseau de publicité Display est constitué d'un ensemble de sites Web qui s'associent aux moteurs de recherche afin d'offrir des espaces publicitaires sur leurs sites, en échange d'un revenu. On peut y retrouver plusieurs formats d'annonces : annonces textuelles, illustrées, vidéo ou format « rich media ». Ces annonces sont intégrées aux sites Web et ont habituellement du contenu qui est, de quelque façon, relié à l'information qui est présentée. Lorsque l'utilisateur clique sur une de ces annonces, il est dirigé vers un site Web qui fait la promotion du produit ou du service annoncé.

Les annonceurs qui choisissent de diffuser leurs annonces sur le réseau Display ont l'option de choisir les sites exacts sur lesquels ils veulent que leurs annonces soient affichées, cibler leurs annonces en fonction d'une liste de mots-clés de recherche ou choisir en fonction des types de clientèles qu'ils désirent atteindre. Peu importe la méthode utilisée pour cibler la clientèle, les annonceurs devront payer pour faire afficher leurs annonces. Généralement, la tarification se fait en fonction du nombre de clics, mais il existe également des tarifications en fonction du nombre d'impressions. Cela dépend du moteur de recherche ainsi que des caractéristiques plus

spécifiques de la publicité. Les tarifs dépendent également de la qualité du site sur lequel l'annonce sera placée; les sites qui génèrent plus d'opportunités de vente auront tendance à coûter plus cher.

Les publicités sur le réseau Search

Les publicités sur le réseau Search représentent le second type de publicité associé aux moteurs de recherche. Elles peuvent être observées sur les pages de résultats, après avoir effectué une requête. Contrairement aux annonces sur le réseau Display, les annonces sur le réseau Search ne peuvent contenir aucune image ni vidéo; elles sont essentiellement constituées de quatre lignes de texte qui présentent le produit ou service offert, ainsi que le site Web qui y est associé (cela explique pourquoi ces publicités sont communément appelées « annonces textuelles »). En cliquant sur ces publicités, l'utilisateur est redirigé vers le site Web de l'annonceur concerné.

Les annonces textuelles ne se retrouvent que sur les pages de recherche; on ne les observe jamais sur des sites Web. De plus, leur tarification est déterminée de façon très différente à celle des publicités Display. Les moteurs de recherche utilisent un système d'enchères fermées selon lequel les annonceurs qui misent le plus augmentent leurs chances d'être bien positionnés dans la page de résultats. Puisque les annonceurs peuvent choisir précisément à quels mots-clés ils désirent associer leurs publicités et peuvent faire varier leurs enchères d'un mot-clé à l'autre, plusieurs considèrent que les publicités du réseau Search sont beaucoup plus efficaces que les publicités sur le réseau Display pour cibler une clientèle précise.

Ce sont uniquement les publicités sous forme d'annonces textuelles (réseau Search) qui ont été étudiées dans le cadre de ce mémoire. En effet, les publicités du réseau Display ne sont pas du tout abordées. Ainsi, la section 1.2 présente, en détail, toutes les explications nécessaires à la compréhension du fonctionnement des mécanismes de publicité sur le réseau Search.

## 1.2 Définitions importantes

Cette section présente plusieurs définitions, explications et exemples qui permettent de comprendre les nombreux termes et expressions propres à l'industrie des annonces textuelles sur

le réseau Search des moteurs de recherche. En présentant les nombreux éléments qui caractérisent les annonces textuelles ainsi que la façon dont ces éléments interagissent ensemble, il est possible de fournir une bonne compréhension du milieu. En effet, les détails présentés dans cette section sont indispensables à la compréhension des concepts qui sont abordés dans le reste du mémoire.

Pour faciliter la compréhension, les termes sont présentés dans un ordre logique plutôt qu'en ordre alphabétique. Ils sont accompagnés de leur traduction en anglais, insérée entre parenthèses, afin de permettre la correspondance avec le vocabulaire que l'on retrouve dans la littérature. Il est important de mentionner que les explications qui suivent caractérisent les mécanismes de publicité de Google, Yahoo et Bing, mais que le fonctionnement peut être différent dans le cas des autres moteurs de recherche.

**Moteur de recherche (« Search engine »)**

Un moteur de recherche est un « outil informatique qui, à partir de critères alphanumériques définis par l'utilisateur, fournit une liste de liens Internet répondant à ces critères » (Microsoft Encarta, 2011). Plus précisément, un moteur de recherche permet de repérer et d'accéder à des pages Web, images, vidéos et autres types de fichiers correspondant à des critères de recherche spécifiques. Les moteurs de recherche les plus efficaces sont ceux qui réussissent à fournir, en peu de temps, les résultats associés à une requête donnée, classés en ordre décroissant de pertinence.

Il existe actuellement quatre grands moteurs de recherche qui possèdent plus de 98% des parts du marché mondial : Google (84,8%), Yahoo (5,7%), Bing (3,9%) et Baidu (3,8%) (netmarketshare.com, 2011). Chacun possède ses propres particularités au niveau de ses algorithmes de classement ainsi que son mécanisme de diffusion des annonces textuelles, mais les principes de base sont très semblables d'un moteur de recherche à l'autre.

Depuis octobre 2010, Yahoo et Bing ont officiellement été fusionnés en un seul moteur de recherche. Les deux sites web sont toujours accessibles indépendamment, mais ils partagent les mêmes algorithmes de recherche. Les résultats obtenus sur ceux-ci seront donc très semblables (ou même identiques) d'une requête à l'autre. Devant ce fait, il est important de mentionner que

dans les analyses qui sont présentées au cours de ce mémoire, les données des deux moteurs de recherche ont été considérées séparément puisqu'elles dataient d'avant la fusion.

Par ailleurs, même si le moteur de recherche Baidu possède 3,3% des parts du marché mondial, il est très rarement utilisé ailleurs qu'en Chine. Ainsi, de notre point de vue, il n'est pas très utile de l'étudier pour des fins commerciales (cela serait même impossible, étant donné l'absence de données historiques). Dans le cadre de cette étude, il ne sera donc pas considéré.

Bref, pour la suite du mémoire, lorsqu'on fait référence aux « moteurs de recherche », on parle de Google, Yahoo et Bing. Ce sont les trois moteurs de recherche qui ont été considérés lors des analyses à effectuer. Les études traitant de Google sont particulièrement intéressantes, étant donnée sa part de marché beaucoup plus grande.

**Utilisateur (« User »)**

Le terme utilisateur est utilisé pour désigner la personne qui effectue des requêtes sur le moteur de recherche dans le but d'obtenir des informations, des biens ou des services.

**Annonceur (« Advertiser »)**

L'annonceur est la personne ou l'entreprise qui crée et gère des campagnes d'annonces textuelles dans le but de faire la promotion de produits ou de services.

**Requête (« Query »)**

Une requête est effectuée lorsqu'un utilisateur exécute une recherche sur une page Web à l'aide d'un moteur de recherche. La requête correspond au terme ou à l'expression qui a été inscrite dans la case de recherche.

**Mot-clé (« Keyword »)**

Un mot-clé est une expression, constituée d'un ou plusieurs mots, à laquelle les annonceurs intéressés peuvent associer une valeur d'enchère qui leur convient (pour une définition de « valeur d'enchère », voir plus loin). Plus la valeur d'enchère est élevée, plus l'annonceur augmente sa probabilité d'obtenir une bonne position dans la liste d'annonces lorsqu'une requête impliquant ce mot-clé sera effectuée. Il existe une infinité de mots-clés potentiels, puisque n'importe quelle expression imaginable pourrait constituer un mot-clé si un annonceur décidait d'y associer une valeur d'enchère.

**Type de correspondance (« Match type »)**

Le type de correspondance définit le degré de similarité qui doit exister entre un mot-clé et une requête pour que l'annonce du mot-clé apparaisse au moment où cette requête est effectuée. En effet, il n'est pas toujours nécessaire que le mot-clé soit identique à l'expression saisie dans la requête pour que l'annonce de ce mot-clé apparaisse sur la page de recherche.

Les types de correspondance peuvent varier légèrement d'un moteur de recherche à l'autre, mais Google, Yahoo et Bing utilisent sensiblement les mêmes, à quelques différences près. Voici les principaux types de correspondance qui sont généralement offerts aux annonceurs :

Mot-clé exact

La requête doit être identique au mot-clé.

Par exemple, si on considère le mot-clé « voiture à vendre » en type « mot-clé exact », il faudra absolument que la requête « voiture à vendre » soit effectuée pour que l'annonce apparaisse.

Expression exacte

La requête doit contenir tous les mots qui constituent le mot-clé, dans l'ordre, mais peut ajouter d'autres mots avant ou après.

Par exemple, si on considère le mot-clé « voiture à vendre » en type « expression exacte », la requête « voiture à vendre pas chère » ferait en sorte que l'annonce du mot-clé apparaîtra.

Requête large

La requête doit contenir, dans l'ordre ou non, un ou plusieurs des mots qui constituent le mot-clé. Ceux-ci peuvent être accompagnés d'autres mots. On accepte également de légères variations des mots telles que le singulier/pluriel, les synonymes ainsi que d'autres variations pertinentes.

Par exemple, si on considère le mot-clé « voiture à vendre » en type « requête large », la requête « jolie voiture » ferait en sorte que l'annonce du mot-clé apparaîtra.

Pour complémenter ces types de correspondance, les moteurs de recherche offrent souvent des paramètres additionnels qui peuvent permettre de mieux cibler certains marchés. Voici les deux plus importants paramètres :

Expression négative

L'utilisation de ce paramètre permet de s'assurer que l'annonce n'apparaît pas lorsque la requête contient l'expression mentionnée.

Par exemple, si on considère le mot-clé « voiture à vendre » avec l'expression « usagée » en type « expression négative », la requête « voiture usagée à vendre » ne fera jamais apparaître l'annonce du mot-clé.

Requête large modifiée (disponible seulement avec Google)

L'utilisation de ce paramètre permet de faire un compromis entre les types de correspondance « expression exacte » et « requête large ». Plus spécifiquement, il permet d'imposer, dans le cas d'un mot-clé de type « requête large », la présence de un ou plusieurs des termes qui la constituent.

Par exemple, si on considère le mot-clé « voiture à vendre » en type « requête large » avec le terme « vendre » précisé comme terme à inclure obligatoirement, toutes les requêtes qui feront apparaître l'annonce de ce mot-clé devront contenir le terme « vendre ». Ainsi, il serait impossible que la requête « jolie voiture » fasse apparaître l'annonce du mot-clé.

Pour chacun des mots-clés qui constituent sa campagne, l'annonceur doit choisir un type de correspondance et il est libre d'y associer un ou plusieurs paramètres additionnels. Les types de correspondance choisis peuvent être différents d'un mot-clé à l'autre, au sein d'une même campagne. Les types « mot-clé exact » et « expression exacte », ainsi que tous les paramètres additionnels sont utilisés par les gestionnaires de campagnes dans le but de mieux cibler, en fonction de la sémantique des mots-clés, le marché auquel ils désirent présenter leurs annonces. Le type « requête large » permet d'éviter l'augmentation exponentielle du nombre de mots-clés dans une campagne en élargissant le marché ciblé par un mot-clé.

Il est possible d'associer plus d'un type de correspondance à un même mot-clé. Ceci constitue une pratique souvent utilisée par les annonceurs, car les coûts par clic des mots-clés en type « mot-clé exact » sont généralement plus faibles que ceux en type « expression exacte », qui sont à leur tour plus faibles que ceux en type « requête large ». En effet, il y a moins de compétition pour les mots-clés avec des types de correspondance plus spécifiques, donc les valeurs d'enchère ont tendance à être plus faibles. Bref, les types de correspondance facilitent la gestion des campagnes et, lorsque utilisés intelligemment, peuvent en améliorer le rendement.

**Campagne (« Campaign »)**

Une campagne est habituellement constituée de plusieurs milliers de mots-clés et vise à faire la promotion d'un ensemble de produits ou de services. 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.

**Annonce textuelle (« Text ad »)**

Une annonce textuelle est essentiellement un texte clair et concis, d'une longueur de 4 lignes au total, qui vise à inciter un clic de la part de l'utilisateur. La première ligne est le titre de l'annonce et constitue le lien à cliquer. Les deuxième et troisième lignes présentent une description du produit ou service offert (ces deux lignes de texte sont communément appelées « le créatif »). Finalement, la dernière ligne indique le site Web de l'annonceur. Parfois, cette adresse URL correspond au site Web sur lequel l'utilisateur sera dirigé, s'il décide de cliquer sur le titre de

l'annonce (ce site est appelé la « page de destination »). Cependant, il arrive souvent que l'annonceur indique plus simplement l'URL de la page d'accueil de son site, même si la page de destination n'est pas la même.


Figure 1.1 : Exemple d'annonce textuelle

**Groupe d'annonces (« Ad group »)**

Une campagne est habituellement constituée de plusieurs groupes d'annonces auxquels il est possible d'associer différents créatifs et différents mots-clés. L'utilité de séparer les mots-clés en plusieurs groupes d'annonces distincts provient du fait qu'il est possible de créer plusieurs annonces textuelles différentes et ainsi mieux cibler les différentes clientèles.

**Impression (« Impression »)**

Un mot-clé obtient une impression lorsqu'une annonce textuelle apparaît sur la page Web d'un moteur de recherche, suite à une requête effectuée par un utilisateur. Au moment où une requête est effectuée, tous les annonceurs ayant fixé une valeur d'enchère suffisamment élevée pour le mot-clé en question verront leur annonce textuelle apparaître sur la page de recherche (pour une définition de « valeur d'enchère », voir plus loin).

**Clic (« Click »)**

Un mot-clé obtient un clic lorsqu'un utilisateur clique sur le titre de l'annonce qui lui est associée. Suite à un clic, l'utilisateur est immédiatement dirigé vers la page Web de destination de l'annonce en question.

**Taux de clic (« Click-through rate » ou « CTR »)**

Le taux de clic d'un mot-clé est calculé en divisant son nombre total de clics par son nombre total d'impressions.

**Position (« Position »)**

La numérotation des positions

La position est une valeur numérique entière qui représente l'ordre dans lequel les annonces textuelles des annonceurs apparaissent sur la page Web du moteur de recherche, une fois qu'une requête est effectuée. La position 1 est toujours la plus haute dans la liste. Plus la valeur de la position augmente, plus l'annonce est située près du bas de la liste. Dans tout le document présent, nous référons aux positions « élevées » comme étant les positions les plus hautes dans la liste (valeurs se rapprochant de 1) et les positions « basses » comme étant celles qui se retrouvent plus près du bas de la liste. Par exemple, dans la Figure 1.2, on observe cinq annonces textuelles distinctes, dans la portion de droite de la page de recherche. Dans ce cas, la position 1 est la plus élevée dans la liste et la position 5 est la plus basse.

Les positions « Premium »

Dans l'exemple illustré à la Figure 1.3, on constate que les trois premières positions sont situées en-haut de la page, directement au-dessus des résultats de recherche. Ces positions sont appelées les positions « Premium » et seront numérotées de 1 à 3. Les moteurs de recherche attribuent parfois ce type de position aux annonces les plus performantes et pertinentes, dans les cas de mots-clés qui génèrent beaucoup de volume. Étant donné leur grande visibilité, ces positions fournissent habituellement des taux de clic beaucoup plus élevés que ceux qui sont obtenus dans les positions de la portion de droite de la page. Les moteurs de recherche n'attribuent pas toujours le même nombre de positions Premium pour une requête donnée et il est même très fréquent qu'une requête ne génère aucune position Premium. Les cinq autres annonces positionnées dans la portion de droite de la page seront numérotées de 4 à 8.

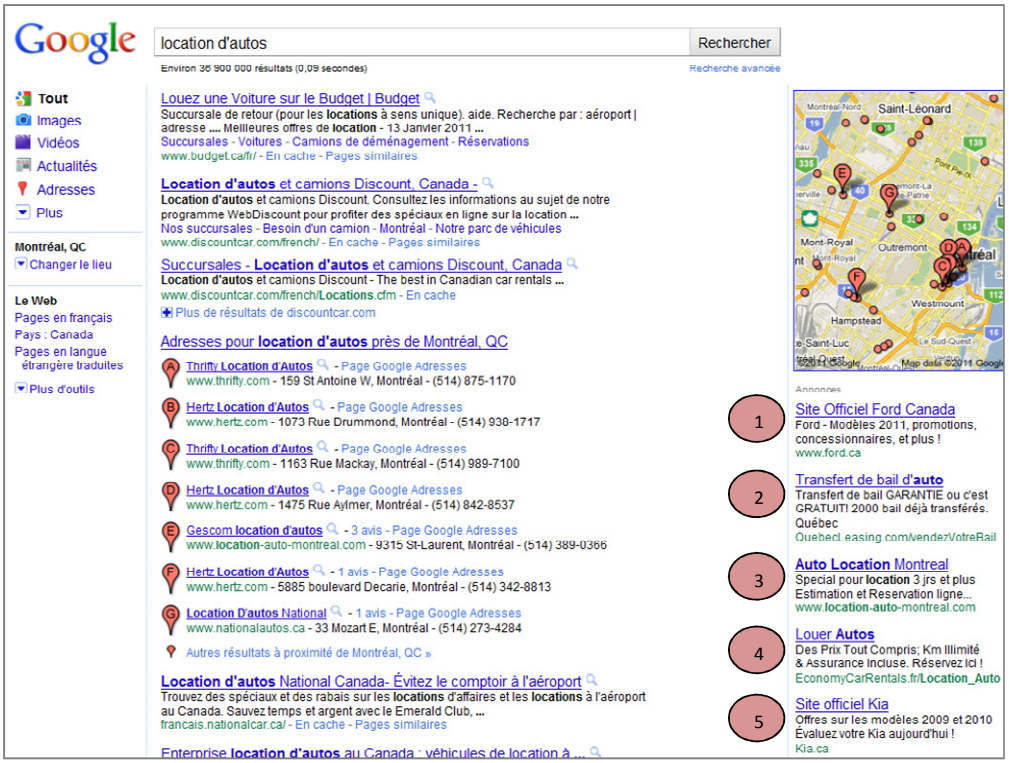



Figure 1.2 : Exemple des positions attribuées aux annonces par Google

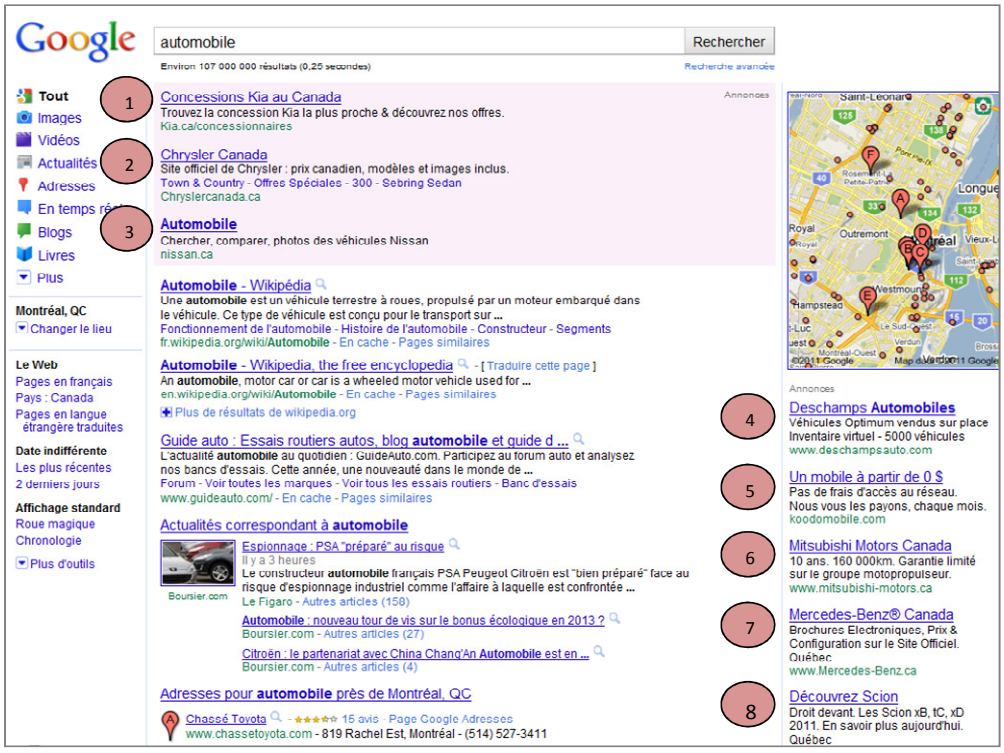



Figure 1.3 : Exemple des positions attribuées aux annonces par Google (cas avec positions Premium)

Il est important de préciser que dans les données historiques fournies par les moteurs de recherche, on ne fait pas la distinction entre les positions Premium et les positions dans la portion de droite de la page de recherche. Par exemple, un mot-clé positionné dans la première position Premium aura une valeur de position de 1, tout comme un mot-clé positionné dans la première position à droite (en l'absence de positions Premium) aura également une valeur de position de 1. Ceci pose quelques problèmes au niveau de l'analyse des données, puisqu'il n'est pas possible de faire la distinction entre les deux types de position. Cependant, puisqu'il est impossible de faire autrement, les analyses ont dû être faites malgré ce manque d'information, en ne faisant pas la distinction entre les positions Premium et les positions standard.

Nombre de positions par page

Le nombre minimal d'annonces affichées sur une page de recherche dépend simplement du nombre de compétiteurs qui ont misé sur ce mot-clé. Pour ce qui est du nombre maximal d'annonces par page, celui-ci peut varier d'une requête à l'autre. Les moteurs de recherche affichent généralement un maximum de 8 à 11 annonces sur une page de recherche. La valeur exacte du nombre maximal d'annonces par page dépendra à la fois du moteur de recherche et du nombre de résultats apparaissant dans les positions de type « Premium ».

S'il existe un grand nombre d'annonces textuelles associées à une requête donnée, il peut être nécessaire de visiter les pages de recherche subséquentes pour voir les annonces qui suivent. Puisqu'il faut que l'utilisateur clique sur le lien « page suivante » au bas de la page de recherche afin de visionner ces annonces, le trafic est généralement beaucoup plus faible que celui de la première page.

Calcul de la position moyenne

Les moteurs de recherche fournissent, sur une base quotidienne, les statistiques relatives au rendement des annonces de leurs clients. Ainsi, les valeurs des positions occupées par un mot-clé sont données sous forme de « position moyenne par jour ». Il s'agit simplement de la moyenne de chacune des positions visitées par l'annonce associée au mot-clé pendant la journée, pondérée par le nombre d'impressions obtenues en chacune de ces positions. Cette valeur de position moyenne peut donc être une valeur non entière.

**Valeur d'enchère ou CPC max (« Bid » ou « Max CPC »)**

Le positionnement des annonces textuelles sur la page de recherche fonctionne selon un mécanisme d'enchères; plus un annonceur fixe une valeur d'enchère élevée, plus il risque de retrouver son annonce dans les positions favorables sur la page de recherche. Ainsi, la valeur d'enchère détermine l'ordre dans lequel les annonces textuelles seront présentées, lorsqu'une requête sera effectuée. Il s'agit d'une valeur qui est choisie par l'annonceur et qui peut être modifiée à n'importe quel moment au cours de la vie d'une campagne. Cette valeur représente le montant maximal que l'annonceur est prêt à payer pour un clic obtenu suite à une requête donnée. L'annonceur doit fixer une valeur d'enchère pour chacun des mots-clés qui constituent sa campagne. Le moteur de recherche ne facturera jamais un coût par clic supérieur à celui qui est précisé par la valeur d'enchère.

**Indice de qualité (« Quality score »)**

Les plus grands moteurs de recherche utilisent un indicateur du niveau de qualité d'un mot-clé pour pondérer la valeur d'enchère au moment du calcul du classement des annonces textuelles. Avec Google, ces valeurs sont connues par les annonceurs. Cependant, avec Bing et Yahoo, les annonceurs ne connaissent pas les valeurs d'indice de qualité qui sont associées à leurs annonces.

Les méthodes utilisées pour calculer les indices de qualité varient d'un moteur de recherche à l'autre, mais elles tiennent compte essentiellement de facteurs de pertinence tels que le taux de clic du mot-clé, la pertinence de la relation entre le mot-clé et l'annonce, ainsi que la pertinence de la relation entre le mot-clé et la page de destination. Cela permet de favoriser les annonces les plus pertinentes, ce qui est avantageux autant pour l'utilisateur que pour l'annonceur et le moteur de recherche.

**Coût par clic (« Cost per click » ou « CPC »)**

Le CPC indique le montant que l'annonceur paie réellement pour chaque clic que son annonce génère. Le CPC facturé à l'annonceur est toujours inférieur ou égal à la valeur d'enchère qu'il a préalablement fixée. En effet, selon l'algorithme « Generalized Second-Price » utilisé pour classer les annonces (expliqué en détail à la section 1.5), les annonces sont toujours classées en

ordre décroissant de leur produit *valeur d'enchère*indice de qualité*. Le CPC facturé à l'annonceur correspond à la valeur minimale nécessaire pour que l'annonce demeure plus élevée que celle du prochain compétiteur dans le classement.

Tout comme les valeurs de position moyenne, les moteurs de recherche fournissent des statistiques quotidiennes concernant les CPC moyens par jour. Puisque les données sont fournies sous un format agrégé, il n'est pas possible de connaître exactement le montant qui a été payé pour chaque clic individuellement. On ne peut que connaître la moyenne des coûts par clic de la journée, qui est calculée en divisant la somme des coûts par la somme des clics obtenus durant la journée.

**Budget quotidien (« Daily budget »)**

Le budget quotidien correspond au montant maximal que l'annonceur souhaite consacrer par jour à sa campagne de publicité. Il faut fixer un budget quotidien pour chacune des campagnes. Dès que la somme des coûts associés aux clics d'une campagne excède le budget quotidien qui lui a été accordé, le moteur de recherche met fin à la diffusion des annonces pour la journée.

**Conversion (« Conversion »)**

Une conversion est comptabilisée pour un mot-clé lorsqu'un utilisateur, dirigé vers le site Web suite à un clic, effectue une action prédéfinie (e.g. achat, inscription, visionnement, demande d'information, etc.). Puisque les conditions nécessaires à l'obtention d'une conversion sont définies par l'annonceur, celles-ci peuvent avoir lieu à des fréquences très différentes d'une campagne à l'autre. Seul l'annonceur peut connaître réellement la valeur monétaire qu'il associe à une conversion.

Pour plusieurs annonceurs, les ventes se font en magasin plutôt que sur le Web. Les liens publicitaires sont simplement utilisés pour faire connaître la marque et les produits, mais les « conversions » ont ultimement lieu ailleurs que sur Internet. Dans ces cas, les statistiques sur les conversions ne sont pas comptabilisées.

**Taux de conversion (« Conversion rate »)**

Le taux de conversion d'un mot-clé est calculé en divisant son nombre total de conversions par son nombre total de clics.

**Enchère minimale en première page (« First page bid »)**

Les moteurs de recherche définissent une valeur d'enchère minimale devant être payée pour apparaître dans la première page des résultats de recherche. Cette valeur peut varier d'un annonceur à l'autre (puisque les annonceurs n'ont pas tous le même indice de qualité) et peuvent aussi varier d'un mot-clé à l'autre (puisque les mots-clés ne sont pas tous autant en demande). Pour qu'une annonce apparaisse dans la première page de résultats de recherche, il est nécessaire que sa valeur d'enchère soit supérieure ou égale à la valeur du « First page bid ».

## 1.3 Historique des méthodes de classement des annonces

Depuis l'apparition des annonces textuelles sur le Web, de nombreuses méthodes différentes ont été employées pour classer les annonces qui s'y retrouvent. Edelman et al. (2006) ainsi que Liu, Chen & Whinston (2008) fournissent une présentation détaillée des différents algorithmes qui ont été utilisés au fil des années. Dans la section qui suit, une brève synthèse de l'histoire de l'évolution des algorithmes de classement est présentée.

Dès 1994, des annonces sur les moteurs de recherche étaient vendues en fonction d'un nombre d'impressions. Ainsi, les annonceurs négociaient des tarifs fixes qu'ils payaient pour faire apparaître leur annonce un certain nombre de fois sur la page de recherche. Cette méthode ne permettait pas aux annonceurs de cibler des clientèles particulières, puisque c'étaient les moteurs de recherche qui déterminaient à quel moment la publicité était présentée aux utilisateurs. De plus, ils imposaient des montants d'achat minimaux qui pouvaient atteindre plusieurs milliers de dollars par mois, des prix que seuls les gros vendeurs pouvaient se permettre. Les petites entreprises ne pouvaient donc pas bénéficier des mécanismes d'annonces textuelles.

Plus tard, en 1997, certains moteurs de recherche (initialement Overture et GoTo) ont introduit des méthodes permettant aux annonceurs de cibler des clientèles particulières, en leur permettant

d'associer leurs annonces à des mots-clés spécifiques. Une valeur d'enchère différente pouvait être associée à chacun des mots-clés et le positionnement des annonces était déterminé en fonction du montant que les annonceurs fixaient comme valeur d'enchère. Plus précisément, les annonces étaient placées en ordre décroissant de leurs valeurs d'enchères, donc ceux qui payaient le plus cher obtenaient plus de visibilité et, par conséquent, plus de chances de générer des clics.

Avec cette méthode, la facturation ne se faisait plus en fonction du nombre d'impressions; les clients étaient facturés pour chaque clic que leur annonce générait. Les annonceurs pouvaient donc déterminer la valeur qu'ils associaient à un clic de chaque mot-clé, en fonction de la pertinence de celui-ci avec le produit ou service qu'ils offraient. Le montant payé par clic correspondait à la valeur d'enchère qui avait été fixée. Avec ce système, les annonceurs bénéficiaient d'une plus grande flexibilité; il était possible de fixer un montant quelconque pour les valeurs d'enchère et on pouvait choisir autant de mots-clés que voulu. Ainsi, autant les grands vendeurs que les petites entreprises pouvaient y afficher leurs publicités.

Après quelque temps, les moteurs de recherche ont constaté que ce mécanisme était instable. En effet, avec ce système, les annonceurs avaient avantage à incrémenter leurs valeurs d'enchère à plusieurs reprises au cours de la journée afin que leurs annonces dépassent celles de leurs compétiteurs. Plus l'annonceur ajustait ses valeurs d'enchère fréquemment au cours de la journée, plus il améliorait ses chances d'obtenir des positions favorables. C'est pourquoi Google a introduit, en 2002, son système de classement basé sur l'algorithme « Generalized Second-Price » (GSP). Dans l'algorithme GSP, les annonces sont classées en ordre décroissant de valeur d'enchère. Cependant, les annonceurs ne paient que la valeur minimale nécessaire pour excéder l'enchère du plus proche compétiteur. Par exemple, si l'annonceur #1 mise 2,00$ par clic et l'annonceur #2 mise 1,00$ par clic, l'annonceur #1 paiera seulement 1,01$ pour chaque clic qu'il recevra. Avec cet algorithme, un équilibre peut être atteint si chacun des annonceurs fixe une valeur d'enchère qui lui convient, car aucun d'entre eux n'aura avantage à changer sa valeur d'enchère en cours de journée pour s'adapter.

Peu après son introduction par Google, l'algorithme GSP fut également adopté par Yahoo. Pour classer ses annonces, Yahoo utilisait une version non modifiée de l'algorithme GSP. Google, pour sa part, utilisait une version légèrement améliorée; les valeurs d'enchère étaient pondérées par les taux de clic des annonces. De cette façon, il était possible de maximiser les revenus

espérés, c'est-à-dire la somme des clics espérés multipliée par le coût par clic. Cette stratégie semblait plus profitable, puisqu'elle tenait compte de la probabilité qu'une annonce soit cliquée.

Aujourd'hui, c'est une autre variante de l'algorithme GSP qui est utilisée pour classer les annonces. Tous les principaux moteurs de recherche utilisent sensiblement le même mécanisme, qui consiste à appliquer l'algorithme GSP en pondérant les valeurs d'enchère par un indice de qualité associé à l'annonce en question. Le calcul de l'indice de qualité varie légèrement d'un moteur de recherche à l'autre, mais il est fortement influencé par le taux de clic des annonces. La seule différence avec la méthode utilisée par Google et Yahoo quelques années plus tôt est le fait qu'on tient compte également de la pertinence de la relation entre le mot-clé et l'annonce, ainsi que la pertinence de la page de destination. Cela fait partie d'une stratégie qu'emploient les moteurs de recherche pour être plus profitables à long terme; plus les utilisateurs ont des expériences positives lors de leurs visites, plus ils auront tendance à revenir y faire des requêtes. C'est pourquoi la prise en compte de la « qualité » de l'annonce est très importante, même si elle ne maximise pas nécessairement les profits à court terme.

## 1.4 L'effet de la position

Puisque les valeurs d'enchère déterminent la position de l'annonce sur la page de recherche, il est important de comprendre comment le nombre de clics générés par une annonce varie en fonction de la position. Pourquoi les positions élevées sont-elles autant en demande? Pourquoi coûtent-elles beaucoup plus cher que les positions plus basses?

### 1.4.1 Les études publiées

Selon nos recherches, il existe un nombre très limité de publications scientifiques traitant de ce sujet. En effet, plusieurs études ont analysé l'effet de la position dans les résultats de recherche non payants (aussi appelés résultats organiques) et ont conclu que le nombre de clics diminuait lorsque la position baissait (notamment Agichtein, Brill, Dumais & Ragno, 2006; Joachims, Granka, Pan, Hembrooke & Gay, 2005; Craswell, Zoeter, Taylor & Ramsey, 2008). Cependant, très peu d'études de ce type ont été effectuées pour analyser l'effet de la position dans le cas des annonces textuelles. La seule étude pertinente effectuée à ce sujet est celle de Agarwal, Hosanagar & Smith (2008). Dans cette étude, les auteurs concluent que la position d'une annonce

a un énorme effet sur la probabilité que cette annonce soit cliquée. Plus spécifiquement, les taux de clic des annonces sont généralement décroissants en fonction de la position.

Même si le sujet n'a pas été abordé fréquemment dans des articles scientifiques, plusieurs entreprises impliquées dans le milieu du marketing sur les moteurs de recherche ont pris l'initiative de procéder à leurs propres analyses, dans le but de mieux comprendre le comportement des mots-clés en fonction de la position. C'est pourquoi il existe, sous forme de documents « White Papers », les résultats de plusieurs études sur ce sujet. Notamment, les études du Atlas Institute (Brooks, 2004a; Brooks, 2004b) et de Enquiro (Hotchkiss, Alston & Edwards, 2005) présentent des résultats intéressants[1]. Brooks (2004a) conclut que les probabilités de clic et le nombre d'impressions sont des variables décroissantes en fonction de la position. Le nombre de clics obtenu dans les positions élevées est donc considérablement plus grand que ce qui est obtenu dans les positions plus basses. De plus, Brooks (2004b) affirme que le potentiel de conversion est également décroissant en fonction de la position. Dans le cas de Hotchkiss, Alston & Edwards (2005), ils utilisent des outils de suivi des yeux pour conclure que la position d'une annonce est fortement corrélée à l'attention qui lui est accordée. Notamment, les positions Premium ont tendance à être observées beaucoup plus souvent que les positions dans la portion de droite de la page de recherche. Par la suite, ils procèdent à des analyses de données pour conclure que les positions les plus élevées ont tendance à générer de meilleurs taux de clic, particulièrement dans le cas des positions Premium.

### 1.4.2 Nos analyses

Pour vérifier l'hypothèse selon laquelle le nombre de clics générés par une annonce dépend fortement de sa position, nous avons choisi d'effectuer nos propres analyses, à partir des données à notre disposition (pour une description détaillée des 20 banques de données utilisées, voir la

[1] Les études effectuées par ces entreprises se retrouvaient dans des « White Papers » et non des articles scientifiques. Cependant, les analyses présentées sont effectuées de façon objective et les méthodes mathématiques utilisées semblent être suffisamment rigoureuses pour que les résultats méritent d'être soulignés.

section 5.2). Les données utilisées totalisent $2{,}45*10^8$ clics et $9{,}45*10^9$ impressions. Elles constituent donc un échantillon de données intéressant pour l'analyse.

L'étude qui a été effectuée est relativement simple; elle consistait à comptabiliser les taux de clic moyens de chacune des banques de données, afin d'évaluer le potentiel associé aux différentes positions. Il semblait préférable d'étudier les taux de clic par position plutôt que le nombre de clics par position, car le taux de clic fournit un meilleur indicateur de la performance d'une annonce. En effet, nous ne pouvons utiliser le nombre de clics à cette fin, car certaines positions sont moins fréquemment visitées et seraient donc injustement pénalisées lors de la comparaison. Puisque le taux de clic tient compte de la fréquence d'apparition de l'annonce, il semblait beaucoup plus logique d'utiliser cette variable comme indicateur de performance pour comparer chacune des positions.

Bien sûr, l'agrégation des taux de clic par position sur plusieurs mots-clés différents implique l'hypothèse que toutes les annonces considérées dans le cadre de l'étude auraient sensiblement la même performance si elles étaient placées dans la même position, lors d'une même requête. Puisque ce n'est pas tout à fait le cas, l'agrégation comporte certains biais. Cependant, elle fournit tout de même une approximation générale du comportement des taux de clic en fonction de la position.

Pour s'assurer d'avoir un nombre significatif de données à chaque position, seules les positions 1 à 10 ont été considérées. À l'Annexe 1, les taux de clic calculés pour chacune des positions de chacune des banques de données sont présentés. De façon générale, les positions les plus élevées obtiennent des rendements considérablement supérieurs, comparativement aux positions plus basses. En effet, une fois toutes les banques de données agrégées, le taux de clic moyen en position 1 est environ 8,4 fois plus élevé que celui en position 10 et les taux de clic moyens décroissent à mesure que la position baisse. La Figure 1.4 illustre la répartition des taux de clic agrégés par position. Pour effectuer ces analyses, les valeurs de position non entières ont été arrondies à la position la plus près et chacun des taux de clic agrégés a été calculé en divisant la somme des clics obtenus par la somme des impressions obtenues à la position donnée.

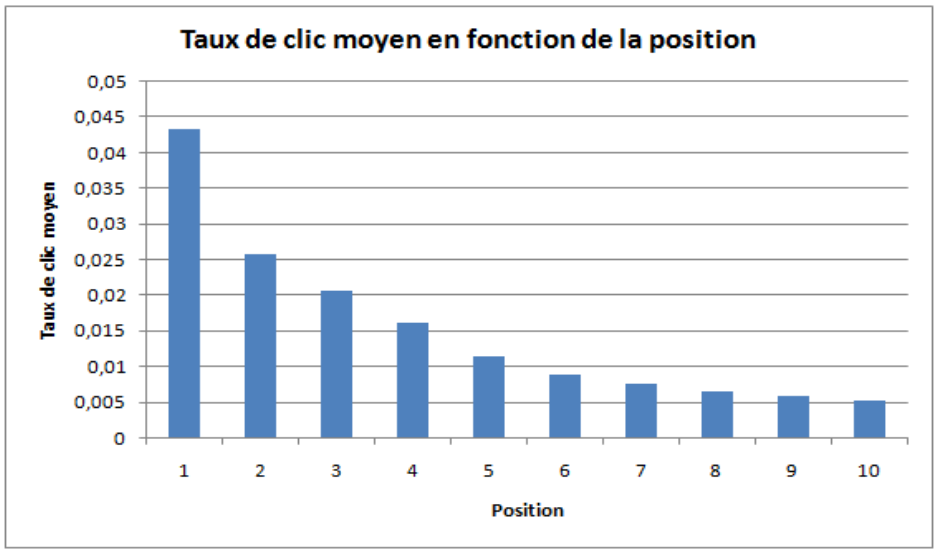



Figure 1.4 : Taux de clic moyens par position (agrégés sur toutes les banques de données)

Finalement, pour éviter les biais mentionnés plus tôt, nous avons effectué les mêmes études que la précédente, mais sur des mots-clés individuels. Un sous-ensemble de plusieurs centaines de mots-clés a été extrait de la base de données afin d'étudier comment le taux de clic moyen variait avec la position. Les mots-clés choisis étaient ceux qui avaient des volumes de clics élevés et bien répartis sur chacune des positions, afin d'observer la variation d'une position à l'autre. De façon générale, les résultats étaient cohérents avec ce qui a été observé au niveau agrégé; les taux de clic semblaient presque toujours diminuer lorsque la position baissait.

### 1.4.3 Les raisons qui expliquent les taux de clic décroissants

Plusieurs facteurs peuvent expliquer le fait que les positions élevées ont tendance à générer plus de clics que les positions basses. Cependant, étant donné le manque d'études traitant de ce sujet, il est impossible de connaître avec certitude les causes de l'effet de la décroissance des taux de clic en fonction de la position. Nous pouvons toutefois poser quelques hypothèses :

1) Les positions plus élevées sont plus fréquemment observées, puisque la plupart des utilisateurs regardent les publicités de haut en bas.

Cette hypothèse est appuyée, entre autres, par Kempe & Mahdian (2008). Ils utilisent un « Cascade model » qui suppose que les annonces sont toujours visionnées de haut en bas. Après avoir observé une annonce, l'utilisateur prend une décision en se basant sur la pertinence de l'annonce visionnée; il peut soit cliquer sur l'annonce ou passer au visionnement de l'annonce suivante dans la liste. Ce modèle suppose donc que la probabilité de clic d'une annonce dépend à la fois de sa position, sa pertinence, ainsi que la pertinence des autres annonces qui la précèdent dans la liste.

2) Les utilisateurs accordent plus de confiance aux résultats plus élevés, car ils considèrent que les moteurs de recherche leur fournissent les résultats en ordre décroissant de pertinence.

Ceci est partiellement vrai, car les moteurs de recherche utilisent effectivement un indice de qualité pour influencer le classement des annonces. Cependant, la valeur d'enchère détermine également l'ordre du classement; une annonce peu pertinente peut donc se retrouver dans les positions élevées si l'annonceur y fixe une enchère suffisamment élevée.

### 1.4.4 Le raisonnement des moteurs de recherche

Nous pouvons supposer que les moteurs de recherche sont arrivés aux mêmes conclusions que celles mentionnées plus haut, puisque la méthode qu'ils utilisent pour facturer les coûts par clic fait en sorte que les positions les plus élevées coûtent plus cher. L'algorithme qu'ils utilisent pour classer les annonces est donc basé sur l'hypothèse que les positions plus élevées ont une plus grande valeur pour les annonceurs. Si ce n'était pas le cas, leur mécanisme de publicité ne fonctionnerait pas, car les annonceurs se contenteraient de payer les prix minimaux pour faire afficher leurs annonces dans les positions inférieures. Il serait illogique de charger plus cher pour quelque chose qui n'a pas plus de valeur.

### 1.4.5 La profitabilité de chaque position

Même si le nombre de clics est presque toujours décroissant en fonction de la position, nous ne pouvons pas affirmer que la profitabilité d'une annonce est toujours maximisée dans les positions

plus élevées. En effet, le profit associé à une annonce est calculé en soustrayant la somme des coûts au total des revenus que cette annonce a généré. Puisque le CPC est décroissant en fonction de la position, les coûts totaux le sont également (considérant que la fonction de clics est décroissante). Ainsi, la profitabilité d'une position dépend du CPC qu'il faut débourser pour l'atteindre. Le problème consiste donc à trouver la position optimale pour laquelle une quantité satisfaisante de clics (et de conversions) est obtenue à un coût acceptable.

À titre d'exemple, observons les données présentées au Tableau 1.1. Ce tableau contient des estimations de rendement relatives à chacune des positions qui pourraient être occupées par un mot-clé quelconque. En supposant que chaque conversion obtenue rapporte un profit moyen de 60,00$ (cette valeur moyenne peut être estimée à partir des données historiques de ventes de l'annonceur) et que le taux de conversion est sensiblement le même pour chacune des positions (cette hypothèse est appuyée, entre autres, par l'économiste en chef chez Google; Varian (2009)), il est possible de calculer les profits nets espérés pour chacune des 10 positions. En observant les valeurs obtenues, nous constatons que les positions les plus élevées ne sont pas les plus profitables, car leurs valeurs de coûts par clic associées sont trop élevées. D'un autre côté, les positions les plus basses sont encore moins profitables, car elles génèrent des quantités de clics beaucoup trop faibles. Dans le cas de ce mot-clé, c'est la position 6 qui fournit le profit le plus élevé. Cette valeur de position optimale peut varier d'un mot-clé à l'autre; elle dépend du nombre de clics qui sera obtenu en chaque position, du coût par clic nécessaire pour atteindre ces positions, des taux de conversions ainsi que des valeurs de profit moyen par conversion.

Tableau 1.1 : Estimations de rendement par position pour un mot-clé quelconque

| Position | Nombre de clics | Coût par clic | Taux de conversion | Nombre de conversions espéré | Profit par conversion | Profit de vente espéré | Coût publicitaire total | Profit net espéré |
|---|---|---|---|---|---|---|---|---|
| 1 | 151 | 2,00 $ | 4,00% | 6,04 | 60,00 $ | 362,40 $ | 302,00 $ | 60,40 $ |
| 2 | 112 | 1,83 $ | 4,00% | 4,48 | 60,00 $ | 268,80 $ | 204,96 $ | 63,84 $ |
| 3 | 80 | 1,50 $ | 4,00% | 3,20 | 60,00 $ | 192,00 $ | 120,00 $ | 72,00 $ |
| 4 | 76 | 1,20 $ | 4,00% | 3,04 | 60,00 $ | 182,40 $ | 91,20 $ | 91,20 $ |
| 5 | 65 | 1,10 $ | 4,00% | 2,60 | 60,00 $ | 156,00 $ | 71,50 $ | 84,50 $ |
| 6 | 60 | 0,84 $ | 4,00% | 2,40 | 60,00 $ | 144,00 $ | 50,40 $ | 93,60 $ |
| 7 | 36 | 0,77 $ | 4,00% | 1,44 | 60,00 $ | 86,40 $ | 27,72 $ | 58,68 $ |
| 8 | 21 | 0,55 $ | 4,00% | 0,84 | 60,00 $ | 50,40 $ | 11,55 $ | 38,85 $ |
| 9 | 12 | 0,43 $ | 4,00% | 0,48 | 60,00 $ | 28,80 $ | 5,16 $ | 23,64 $ |
| 10 | 6 | 0,15 $ | 4,00% | 0,24 | 60,00 $ | 14,40 $ | 0,90 $ | 13,50 $ |

### 1.4.6 Conclusion

La position d'une annonce est une variable importante à considérer dans le contexte d'optimisation du rendement d'une campagne publicitaire. Elle peut influencer considérablement le nombre de clics, le nombre de conversions et le profit net obtenu. Le positionnement des annonces peut faire la différence entre le succès ou l'échec d'une campagne publicitaire.

## 1.5 L'algorithme de classement des annonces

Tous les principaux moteurs de recherche utilisent une variante de l'algorithme GSP pour effectuer le classement de leurs annonces et ensuite déterminer le montant exact qui sera facturé par clic à chacun des annonceurs. Il est donc très important de comprendre le fonctionnement de cet algorithme.

D'abord, au moment où une requête est effectuée, tous les annonceurs ayant misé sur le mot-clé en question sont ordonnés en ordre décroissant de leur produit *valeur d'enchère * indice de qualité.* Par la suite, si une annonce est cliquée, le coût par clic (CPC) qui est facturé à l'annonceur correspond au montant minimal nécessaire pour que celui-ci maintienne son rang dans la liste. Le CPC dépend donc de la valeur d'enchère et de l'indice de qualité de l'annonce à la position subséquente. Plus précisément, le montant facturé à l'annonceur pour un clic est donné par la formule suivante :

$$CPC_p = max\left\{\frac{enchère_{p+1} * IQ_{p+1}}{IQ_p} + \varepsilon\ ;\ CPC_{min}\right\}$$

où

- $p$ est la position à laquelle le clic a été obtenu
- $enchère$ est la valeur d'enchère fixée par l'annonceur sur ce mot-clé
- $IQ$ est l'indice de qualité de l'annonce
- $\varepsilon$ est une valeur minimale d'incrément (habituellement 0,01$)
- $CPC_{min}$ est un coût minimal pour la combinaison annonceur-position

Remarque #1 :

Les moteurs de recherche attribuent des seuils minimaux de coût par clic ($CPC_{min}$), qui varient en fonction des mots-clés et des positions. Il existe donc un prix de réserve pour chacune des positions d'un mot-clé. Ainsi, l'annonceur qui se trouve en dernière position devra tout de même payer un prix pour faire afficher sa publicité.

Remarque #2 :

Le mécanisme que les moteurs de recherche utilisent pour classer les annonces est un mécanisme à « enchères fermées », c'est-à-dire qu'il n'est pas possible de connaître les valeurs d'enchère des compétiteurs. De cette façon, les moteurs de recherche s'assurent que les annonceurs misent le montant maximal qu'ils peuvent se permettre de payer et non le montant minimal nécessaire pour atteindre une position voulue. En effet, puisque les enchères sont fermées, il n'est pas possible de connaître de façon certaine la position qui sera obtenue avec une valeur d'enchère précise.

Exemple

Afin d'illustrer l'application de cet algorithme de classement, considérons l'exemple suivant. Supposons que trois annonceurs différents ont misé sur un mot-clé :

Tableau 1.2 : Valeurs d'enchère et indices de qualité

| | valeur d'enchère | indice de qualité |
|---|---|---|
| Annonceur A | 1,50$ par clic | 5 |
| Annonceur B | 2,00$ par clic | 4 |
| Annonceur C | 1,00$ par clic | 9 |

Nous classons d'abord ces annonceurs en ordre décroissant de leur produit *valeur d'enchère* * *indice de qualité* (ici appelé score).

Tableau 1.3 : Classement des annonceurs

| | valeur d'enchère | indice de qualité | score | classement |
|---|---|---|---|---|
| Annonceur C | 1,00$ par clic | 9 | 9,00 | 1er |
| Annonceur B | 2,00$ par clic | 4 | 8,00 | 2e |
| Annonceur A | 1,50$ par clic | 5 | 7,50 | 3e |

Une fois le classement effectué, calculons le coût par clic qui serait payé par chaque annonceur (supposons $\varepsilon = 0{,}01\$$ et $CPC_{min} = 0{,}05\$$ ) :

$$CPC(C) = max\left\{\frac{2{,}00 \ * \ 4}{9} + 0{,}01 \ ; \ 0{,}05\right\} = 0{,}90 \ \$/clic$$

$$CPC(B) = max\left\{\frac{1{,}50 \ * \ 5}{4} + 0{,}01 \ ; \ 0{,}05\right\} = 1{,}89 \ \$/clic$$

$$CPC(A) = max\left\{\frac{0 \ * \ 0}{5} + 0{,}01 \ ; \ 0{,}05\right\} = 0{,}05 \ \$/clic$$

Il est intéressant de noter que l'annonceur B est fortement pénalisé parce que son indice de qualité est très faible. Par conséquent, il doit débourser beaucoup plus que ses compétiteurs pour être en position 2. Avec un tel mécanisme de classement, les annonceurs ont avantage à choisir des mots-clés qui sont pertinents avec leurs annonces et sites Web.

## 1.6 La gestion d'une campagne

### 1.6.1 Les objectifs

Ce sont les conversions qui génèrent, directement ou indirectement, les revenus des annonceurs. Ainsi, à la base, presque tous les annonceurs ont comme objectif de maximiser leur nombre de conversions en respectant un budget fixe. Cependant, puisque le nombre de conversions peut parfois être très faible, plusieurs d'entre eux choisissent une option quasi-équivalente; ils visent à maximiser le nombre de clics obtenu en respectant une contrainte de budget fixe. En supposant que les probabilités de conversion sont égales d'un mot-clé à l'autre et d'une position à l'autre, cela équivaut à maximiser les conversions.

Le fait d'utiliser le nombre de clics comme indicateur de performance d'une campagne peut être intéressant, car les clics sont beaucoup plus abondants que les conversions et ils permettent donc de quantifier la performance de chacun des mots-clés individuellement. De plus, l'utilisation des conversions comme indicateur de performance mènerait rapidement à la conclusion que la forte majorité des mots-clés qui constituent une campagne publicitaire possèdent un nombre de conversions nul. Par conséquent, leur rendement ne pourrait pas être quantifié et comparé.

Certains annonceurs ont parfois des objectifs un peu différents. En effet, ceux qui n'offrent ni produits ni services cherchent parfois à maximiser le nombre d'impressions obtenues par leur annonce, afin de donner de la visibilité à leur marque et ainsi accroître leur popularité. Cependant, le nombre d'annonceurs avec de tels objectifs est assez limité. Bref, les variables à optimiser varient en fonction des besoins de chaque client, mais ce sont habituellement les conversions ou les clics que les annonceurs cherchent à maximiser.

En gardant leurs objectifs en vue, les annonceurs doivent constamment modifier leurs campagnes afin d'optimiser leur rendement. Ils doivent périodiquement réévaluer la pertinence de leurs mots-clés et de leurs annonces, en fonction des données qu'ils recueillent. De plus, la nature dynamique du milieu fait en sorte qu'il faut ajuster les enchères des mots-clés très fréquemment pour assurer un bon retour sur investissement.

### 1.6.2 La gestion des enchères

Tel qu'expliqué précédemment, les deux facteurs qui influencent la position d'une annonce sont la valeur d'enchère et l'indice de qualité. Puisque l'indice de qualité est fortement dépendant du taux de clic d'une annonce, un annonceur peut très difficilement, à court terme, contrôler la valeur de ses indices de qualité. Il doit donc jouer sur ses valeurs d'enchère pour tenter d'atteindre ses objectifs. La gestion des enchères consiste à fixer et ajuster les valeurs d'enchère de chacun des mots-clés d'une campagne, dans le but de maximiser le rendement global de la campagne.

## 1.7 L'importance des mathématiques

Le monde des annonces textuelles sur les moteurs de recherche est extrêmement complexe et le comportement des consommateurs en ligne est très difficile à prédire. À cause de la taille des campagnes, la périodicité de la demande ainsi que la nature dynamique et compétitive du milieu, il est quasi-impossible pour un être humain de bien gérer une campagne sans avoir recours à des outils d'analyse mathématique automatisés. Le nombre maximal théorique de mots-clés pouvant

constituer une campagne publicitaire est potentiellement illimité et pour chacun de ces mots-clés, il faut réussir à fixer une valeur d'enchère adéquate.

La présence de nombreuses variables et contraintes à respecter en fait un domaine parfait pour l'application de techniques de prédiction, de modélisation et d'optimisation mathématiques. Puisque la complexité d'une campagne croît de façon exponentielle en fonction du nombre de mots-clés qu'elle comporte, les algorithmes d'optimisation automatisés constituent un outil indispensable pour tout annonceur qui désire gérer ses campagnes de façon efficace. Des modèles de prédiction doivent préalablement être élaborés afin de prédire le comportement des diverses variables en jeu; des variables telles que le nombre d'impressions, le nombre de clics, le nombre de conversions et les valeurs d'enchère pour chaque position potentielle doivent être étudiés et modélisés afin de fournir des données pertinentes et fiables aux modules d'optimisation. Par la suite, des méthodes d'optimisation efficaces et robustes doivent être appliquées à ces données, afin de fournir une aide à la décision pour la gestion des enchères des campagnes. Bref, le monde des annonces textuelles représente un environnement idéal pour l'application de diverses méthodes mathématiques.

## 1.8 Les intérêts de chacun des joueurs

Les différents joueurs impliqués dans le monde des annonces textuelles sur les moteurs de recherche sont les annonceurs, les utilisateurs et le moteur de recherche lui-même. Chacun de ces groupes possède des objectifs spécifiques et c'est grâce à un mécanisme de publicité très élaboré et sophistiqué que les intérêts de chacun sont pris en considération.

D'abord, le moteur de recherche vise à maximiser ses revenus de publicité. Ces revenus étant payés par les annonceurs à chaque fois qu'une annonce textuelle est cliquée, il souhaite donc maximiser le nombre de fois que les utilisateurs cliquent sur les annonces qui leur sont présentées. Dans le but de maintenir une rentabilité à long terme, les principaux moteurs de recherche tels que Google, Yahoo et Bing ont choisi d'intégrer des algorithmes de calcul de la pertinence des annonces dans leur mécanisme de classement. Ainsi, les annonces les plus pertinentes sont priorisées au moment du classement, ce qui fait en sorte que les utilisateurs sont plus souvent satisfaits des résultats de leurs recherches et n'hésiteront pas à utiliser ce moteur de

recherche pour effectuer leurs requêtes dans le futur. De plus, puisque l'indice de qualité est fortement dépendant de la probabilité de clic d'une annonce, le moteur de recherche réussit à maximiser ses revenus espérés en classant les annonces par produit décroissant de *valeur d'enchère*indice de qualité*.

Du point de vue des annonceurs, la majorité des campagnes ont comme objectif principal de maximiser le nombre de conversions avec un budget fixe ou de maximiser le nombre de clics avec un budget fixe. Dans cette optique, le fait de payer pour chaque clic représente une opportunité très intéressante. En effet, contrairement aux mécanismes de publicité traditionnels tels que les annonces à la télévision, à la radio ou sur des affiches, l'enchère de coût par clic avec les annonces textuelles assure aux annonceurs qu'ils n'auront à payer des frais que si un client a été redirigé vers leur site Web. Ainsi, les coûts sont directement proportionnels au nombre d'utilisateurs qui réagissent positivement à la publicité.

De plus, le mécanisme d'enchère permet aux annonceurs de choisir le montant maximal qu'ils sont prêts à payer pour chaque clic, en fixant une valeur de max CPC. L'annonceur ne paie que le montant minimal nécessaire pour demeurer une position au-dessus de son plus proche compétiteur, ce qui fait en sorte qu'il n'a jamais le sentiment d'avoir payé plus que nécessaire pour un clic. Le fait que la valeur d'enchère puisse être contrôlée par l'annonceur fait en sorte qu'il est possible pour celui-ci de prendre des décisions en fonction de calculs statistiques reliés à son historique de données. Le nombre de clics et le coût par clic étant généralement décroissants à mesure que la position d'une annonce devient plus basse, il suffit de déterminer le compromis idéal entre le volume de clics et les dépenses associées à ces clics pour décider de la valeur d'enchère à fixer. En effet, en calculant, à partir de l'historique de données, le taux de conversion espéré et le revenu moyen associé à une conversion, il est possible pour les annonceurs de déterminer le revenu espéré qu'ils associent à un clic. Pour avoir une campagne profitable, il suffit de fixer une valeur d'enchère inférieure à ce seuil; lorsque les revenus espérés sont supérieurs aux coûts, la profitabilité à long terme est presque assurée.

Du point de vue des utilisateurs, les algorithmes de classement des annonces font en sorte que les annonces les plus pertinentes sont fortement priorisées. Ainsi, les annonces qui apparaissent dans les positions élevées sur la première page de recherche seront celles qui, selon l'algorithme du moteur de recherche, auront la plus grande probabilité de plaire à l'utilisateur. Il s'agit

d'annonces qui sont habituellement étroitement reliées à la requête et qui mènent vers des pages de destination pertinentes.

Finalement, la nature des enchères de type coût par clic fait en sorte qu'il est possible d'utiliser le sens des mots-clés contenus dans les requêtes pour mieux cibler le marché et même adapter ses annonces en fonction des requêtes. En effet, il est possible de définir des annonces différentes pour plusieurs groupes de mots-clés. Des options de ciblage géographique existent également afin de permettre aux annonceurs d'atteindre uniquement les gens concernés par leurs publicités. Pour chacune des campagnes, il est possible de préciser les régions géographiques dans lesquelles ils désirent faire afficher leurs annonces. Toutes ces possibilités font en sorte que les annonceurs qui gèrent bien leurs campagnes réussissent habituellement à rejoindre leur clientèle cible. Par conséquent, les utilisateurs bénéficient également de la situation, en retrouvant une plus grande proportion d'annonces pertinentes relatives à leurs besoins lorsqu'ils effectuent une requête. Bref, le mécanisme utilisé pour classer les annonces textuelles, les faire afficher et facturer les coûts considère les besoins et intérêts de chacun des joueurs impliqués.

## 1.9 Présentation de l'entreprise

### 1.9.1 Les agences publicitaires

Depuis quelques années, de nombreuses entreprises ont émergé sur le marché en tant qu'agences spécialisées en gestion de campagnes publicitaires sur les moteurs de recherche. Étant donné la complexité et la taille de certaines campagnes de publicité, leur expertise est très en demande.

En effet, les entreprises qui ont de gros volumes de vente et beaucoup de trafic sur leurs sites Web ont presque toujours recours à des agences publicitaires pour gérer leurs campagnes d'annonces textuelles. Les grandes campagnes étant habituellement très complexes à gérer, il est souvent plus efficace de se tourner vers des experts, même si cela peut engendrer des coûts non négligeables. Les agences se chargent de la gestion et de l'optimisation du rendement des campagnes, donc il peut être rentable à long terme pour les entreprises vendeuses de produits ou de services de leur confier la gestion de leurs campagnes.

### 1.9.2 Les logiciels de gestion de campagnes

Pour gérer la complexité de leurs campagnes, les agences utilisent presque toujours des logiciels de gestion de campagnes. Ces logiciels spécialisés permettent habituellement de créer des campagnes, fixer des valeurs d'enchères, automatiser l'ajustement des valeurs d'enchères à l'aide de règles simples ou d'algorithmes plus complexes, optimiser le rendement en fonction d'objectifs précis et générer des rapports périodiques d'évaluation de la performance.

L'industrie des logiciels de gestion de campagnes publicitaires devient de plus en plus compétitive. Plusieurs entreprises visent à offrir leur produit aux nombreux utilisateurs potentiels sur le marché. Afin de se démarquer, celles-ci tentent d'apporter des modifications et ajouts à leur logiciel dans le but d'améliorer le rendement des campagnes de leurs clients et d'en faciliter la gestion. Entre autres, il est devenu de plus en plus fréquent d'effectuer de la recherche et du développement dans le but d'intégrer des solutions d'analyse statistique et d'optimisation aux logiciels. En exploitant au maximum l'information fournie quotidiennement par les moteurs de recherche, il est possible pour les agences et autres gestionnaires de campagnes de développer des stratégies de marketing efficaces.

### 1.9.3 Acquisio

Acquisio est une entreprise qui développe et commercialise une plateforme logicielle de gestion des campagnes publicitaires sur les moteurs de recherche, sur les réseaux sociaux et avec les mécanismes de type « Ad Exchange ». Cette plateforme est destinée principalement aux agences publicitaires, qui peuvent y accéder sur le Web en payant des frais mensuels ou annuels. Elle permet aux annonceurs de créer, suivre, gérer et optimiser leurs campagnes sur les divers canaux publicitaires à leur disposition. De plus, elle leur fournit les outils nécessaires pour faciliter la création de rapports de performance périodiques. Actuellement, la plateforme logicielle d'Acquisio aide plus de 300 agences interactives à gérer des campagnes publicitaires dont les budgets totalisent plus de 500 millions de dollars. Toutes les études présentées dans le cadre de ce mémoire ont été effectuées en collaboration avec Acquisio et à partir de données provenant des campagnes de leurs clients.

Puisque Acquisio vise principalement le marché des agences publicitaires, sa plateforme logicielle est particulièrement efficace pour la création de rapports périodiques de performance.

En effet, les agences doivent habituellement fournir un suivi mensuel à leurs clients, donc l'habileté de générer rapidement des rapports clairs et bien structurés est un atout. De plus, la plateforme logicielle est également très performante au niveau de la gestion automatisée des enchères à l'aide de règles prédéfinies (appelées communément « bid rules » dans le domaine). Les agences peuvent donc se construire des algorithmes simples, à base de règles, qui répondent à leurs besoins spécifiques.

Depuis quelque temps, Acquisio souhaite développer des algorithmes d'optimisation plus complets qui tiendraient compte de la performance de la campagne dans son ensemble. Les méthodes à base de règles sont efficaces et simples à utiliser, mais elles ne permettent que d'optimiser chaque mot-clé individuellement. De nombreuses recherches sur le sujet nous poussent à croire que l'emploi de méthodes de statistique et d'optimisation rigoureuses, avec une optique plus globale, permettraient d'améliorer davantage le rendement des campagnes. Une telle technologie donnerait à Acquisio l'opportunité de se démarquer de ses concurrents en proposant un outil unique dans le domaine.

## 1.10 Présentation du sujet de recherche

Dans le cadre de ce projet de maîtrise, nous avons travaillé en collaboration avec Acquisio dans le but de définir les besoins de l'entreprise en termes de solutions mathématiques, proposer des méthodes et algorithmes qui répondent aux besoins, puis finalement tester, évaluer et comparer chacune des alternatives proposées. Dans la suite du document, le cheminement de notre projet de recherche est présenté en détail.

D'abord, nous présentons une revue de littérature afin de situer les besoins d'Acquisio par rapport à ce qui a déjà été publié dans le domaine. Suite à de nombreuses recherches, nous constatons qu'il n'existe que très peu de publications scientifiques traitant de ce sujet dans la littérature. La nouveauté du sujet de recherche et la nature compétitive du milieu font en sorte qu'il est difficile de trouver des études bien documentées. Parmi les méthodes publiées, nous constatons que celles-ci sont soit sous-optimales ou inutilisables en pratique. Le besoin de développer de nouveaux algorithmes spécifiques à nos besoins apparaît donc évident.

Par la suite, le monde des annonces textuelles est modélisé et représenté sous forme d'un modèle linéaire de type « sac-à-dos binaire à choix multiple ». Le problème consiste essentiellement à maximiser le nombre de clics générés par la campagne, en respectant, entre autres, une contrainte de budget.

Une fois le modèle d'optimisation linéaire présenté, les données utilisées pour effectuer les tests sur chacune des méthodes algorithmiques envisagées sont présentées et décrites de façon détaillée. De cette manière, les lecteurs peuvent comprendre la nature et les caractéristiques de l'échantillon qui est utilisé pour ultimement arriver à des conclusions générales concernant l'industrie des annonces textuelles.

Suite à la description des données, des méthodes de classification et prétraitement des données sont présentées, expliquées et analysées. La méthode de classification permet d'identifier les mots-clés qui peuvent potentiellement servir d'intrant à notre algorithme d'optimisation, ceux qui nécessiteront un prétraitement avant l'optimisation, ainsi que ceux qui devront être exclus totalement de l'optimisation. Pour sa part, la méthode de prétraitement offre l'opportunité d'acquérir de l'information supplémentaire par rapport à certains mots-clés qui, initialement, n'avaient pas les caractéristiques nécessaires pour accéder directement à l'optimisation.

Finalement, des méthodes de prédiction génériques sont développées dans le but d'augmenter considérablement notre potentiel de prédiction des fonctions de clics et de CPC des mots-clés. Trois méthodes alternatives sont testées, évaluées et comparées dans le but de trouver celle qui offre le plus de flexibilité et minimise les erreurs de prédiction. Afin de faciliter l'utilisation de ces méthodes de prédiction dans un contexte pratique, des recommandations concernant l'intégration des fonctions génériques au modèle d'optimisation sont fournies.

# CHAPITRE 2 REVUE DE LITTÉRATURE

L'optimisation du rendement des campagnes d'annonces textuelles sur les moteurs de recherche est un sujet très peu abordé dans la littérature. En effet, le marché des annonces textuelles n'existe que depuis 1997 et les références les plus anciennes publiées à ce sujet datent du début des années 2000. De plus, la nature compétitive du domaine fait en sorte que les entreprises qui subventionnent des recherches dans ce domaine ne publient que très rarement leurs résultats. La plupart des études publiées jusqu'à présent ont été effectuées dans l'optique d'optimisation des profits pour les moteurs de recherche; l'optimisation des enchères de mots-clés du point de vue des annonceurs est très peu fréquemment publiée.

La constante évolution des algorithmes constituant les moteurs de recherche fait en sorte que plusieurs références deviennent rapidement non pertinentes. En effet, les compagnies telles que Google, Microsoft et Yahoo modifient très fréquemment leurs mécanismes de classement des annonces ainsi que leurs méthodes d'attribution des coûts, donc même des études relativement récentes peuvent parfois devenir désuètes suite à l'apport de tels changements. Par exemple, la transition d'une enchère ouverte à une enchère fermée chez Overture (Yahoo) il y a quelques années a considérablement modifié les stratégies de choix d'enchères du point de vue des annonceurs. Puisque le nombre de positions ainsi que les coûts associés à chaque position n'étaient plus connus, il fallait modifier les algorithmes pour tenir compte de l'incertitude associée à la prédiction des coûts par position.

Jansen et Mullen (2008) fournissent une excellente introduction aux concepts du marketing avec des campagnes d'annonces textuelles. Ils expliquent le concept des enchères de mots-clés, présentent un historique de l'évolution des méthodes de classement utilisées et décrivent de façon détaillée leur fonctionnement. Ils fournissent également certaines explications fondamentales quant aux stratégies permettant de gérer des campagnes de façon efficace. Bref, le document permet de comprendre les concepts qui définissent le problème à l'étude et réussit à mettre en évidence la complexité du domaine.

Edelman, Ostrovsky & Schwarz (2006) expliquent de façon détaillée le fonctionnement et l'évolution des divers algorithmes de classement des annonces et de détermination des coûts par clic. Ils présentent plusieurs exemples qui permettent de comprendre la différence entre chacune

des méthodes et ils étudient leurs propriétés du point de vue de la théorie des jeux. Plus particulièrement, ils abordent l'algorithme « Generalized Second-Price », utilisé actuellement par les principaux moteurs de recherche tels que Google et Bing. Cet algorithme classe les annonces en ordre décroissant de score *valeur d'enchère*indice de qualité* et le prix réellement payé par l'annonceur correspond au montant minimal nécessaire pour excéder le score de l'annonceur qui le suit au classement.

Adjengue, Chan, Gamache, Marcotte & Savard (2008) présentent une étude de faisabilité préliminaire dans laquelle ils étudient l'interaction entre les différentes variables qui caractérisent les mots-clés. Leur étude détermine, entre autres, qu'il est possible d'utiliser des méthodes de régression pour prédire le comportement des mots-clés. Ils mentionnent l'importance de tenir compte des fluctuations temporelles intrinsèques à une campagne publicitaire. De plus, ils procèdent à plusieurs tentatives de regroupement des mots-clés (« clustering ») dans le but de plus facilement gérer les mots-clés à faible volume. Ils finissent par conclure qu'il est concevable qu'un système d'optimisation automatisé maximisant les conversions puisse être développé, à condition que les campagnes concernées soient relativement riches en conversions.

## 2.1 Quelques modélisations intéressantes

Kitts, Laxminarayan, Leblanc & Meech (2005) proposent une modélisation mathématique formelle pour les enchères d'annonces textuelles. Ils énoncent également quelques théorèmes, accompagnés de preuves mathématiques, qui permettent de mieux comprendre les concepts à la base de l'optimisation des valeurs d'enchères. Entre autres, ils discutent de l'équivalence entre la maximisation des revenus sous des contraintes de retour sur investissement minimal et la maximisation des conversions sous des contraintes de coût par conversion maximal, puis proposent des valeurs d'enchères optimales pour chacun de ces cas. Ils expliquent ensuite la distinction entre les stratégies d'optimisation locale et celles d'optimisation globale. Finalement, ils abordent le sujet du comportement à l'équilibre des enchères et étudient les effets de l'ajout ou du retrait d'un compétiteur pour un mot-clé donné. Bref, leur étude vise à fournir une meilleure compréhension de l'aspect mathématique du problème, mais ne propose pas de méthode systématique pour gérer des campagnes de mots-clés de façon optimale.

Hou, Wang & Yang (2008) modélisent les interactions entre les divers éléments du modèle sous forme de réseaux bayésiens qui représentent les relations causales entre la valeur de l'enchère, la position de l'annonce, les impressions et le nombre de clics obtenus. Ils proposent d'abord un modèle statique, qui suppose que les conditions du marché demeurent les mêmes à travers le temps. Ils présentent ensuite un modèle dynamique, plus représentatif de la réalité, qui tient compte de l'évolution des conditions du marché d'une journée à l'autre. Dans ce dernier modèle, les probabilités conditionnelles associées aux variables du réseau bayésien sont calculées à l'aide de méthodes d'apprentissage de paramètres, puis des prédictions dynamiques du nombre de clics sont effectuées en tenant compte de la valeur des variables du problème dans les périodes précédentes.

Suite à des essais expérimentaux, l'étude conclut que le modèle de prédiction dynamique a une performance nettement supérieure à celle du modèle statique au niveau de la précision des prédictions, ce qui n'est pas surprenant considérant que le modèle dynamique tient compte de l'historique des données. L'article est particulièrement intéressant pour les modélisations qui y sont présentées; celles-ci illustrent clairement l'interaction qui existe entre chacune des variables du problème.

## 2.2 Méthodes à valeur d'enchère et position constantes

Certains auteurs ont tenté de simplifier le problème en ignorant les concepts de positions et valeurs d'enchères multiples par mot-clé. Dans ce cas, ils supposent qu'il n'existe qu'une seule position et une seule valeur d'enchère par mot-clé. Le problème consiste alors à sélectionner un sous-ensemble de mots-clés qui maximisera les profits, en respectant une contrainte de budget.

Rusmevichientong & Williamson (2006) utilisent une modélisation de ce type et proposent un algorithme glouton qui ordonne les mots-clés en ordre décroissant de ratio *profit estimé/coût estimé*, puis sélectionne les mots-clés en suivant l'ordre de la liste jusqu'à ce que le budget soit épuisé. Le problème est d'abord étudié dans le cas statique, soit lorsque les probabilités de clic sont connues pour chacun des mots-clés. Ils abordent ensuite le cas dynamique, dans lequel les probabilités de clic des mots-clés ne sont pas connues à l'avance et doivent être estimées au fil du temps. Ce second cas implique de considérer le compromis entre l'exploration de nouveaux

mots-clés, afin d'acquérir davantage d'informations quant à leur profitabilité, et l'exploitation des mots-clés déjà connus comme étant profitables. Les auteurs formulent une stratégie d'enchère adaptative qui exploite la structure du problème et qui permet d'orienter les choix de mots-clés en fonction de l'information acquise à chaque jour.

Muthukrishnan, Pal & Svitkina (2010) utilisent également une modélisation à valeur d'enchère et position constantes par mot-clé, mais ils intègrent la dimension stochastique du problème. Ils proposent des modèles qui tiennent compte des distributions statistiques relatives au nombre d'impressions et au nombre de clics, les deux variables de leur problème. Leur approche combine donc la difficulté de prédire avec précision la valeur de l'objectif en fonction d'un ensemble de valeurs d'enchères et la difficulté d'optimiser l'espérance de cette fonction-objectif en se basant sur les modèles statistiques obtenus.

Par ailleurs, Even Dar, Mirrokni, Muthukrishnan, Mansour & Nadav (2009) étudient plus particulièrement l'optimisation du choix des mots-clés dans le cas des types de correspondance à « requête large ». Tout comme avec les deux méthodes précédentes, ils utilisent une approche qui suppose une seule valeur d'enchère et une seule position par mot-clé. Ils développent deux modèles d'optimisation : un modèle ayant comme objectif de maximiser le profit espéré et un autre modèle visant à maximiser le revenu en imposant une contrainte de budget fixe. De plus, ils présentent des algorithmes pour les résoudre en temps polynomial. Leur approche est unique, car elle tient compte des dépendances entre les requêtes, phénomène non négligeable dans le cas des types de correspondance à requête large. Puisque ce type de correspondance élargit l'ensemble des requêtes pouvant être associées à un mot-clé et que les requêtes n'ont pas toutes la même valeur de revenu espéré par clic, il est intéressant de considérer son effet sur la performance globale d'une campagne publicitaire.

Les méthodes à valeur d'enchère et position fixes permettent d'identifier des sous-ensembles de mots-clés profitables. Cependant, la modélisation du problème avec cette approche peut être considérée comme relativement simpliste et comporte plusieurs limites; il n'est pas réaliste de faire abstraction des concepts de position et de valeur d'enchère variables par mot-clé si nous souhaitons optimiser une campagne dans son ensemble. En effet, la position influence énormément le rendement d'une annonce, tel que démontré par les études mentionnées dans la section 1.4. De plus, les CPC sont considérablement différents d'une position à l'autre. Il est

donc primordial de tenir compte de l'ensemble des positions potentielles et de leurs CPC associés si nous désirons considérer toutes les facettes du problème et fournir des solutions réellement profitables.

## 2.3 Méthodes à CPC et positions multiples

Chakrabarty, Zhou & Lukose (2008) mettent en évidence les différences entre le cas à position unique et celui à positions multiples. Ils modélisent le problème sous forme de sac à dos avec arrivée dynamique des informations. Dans le cas à position unique, ils visent à choisir un sous-ensemble de mots-clés qui maximisera les profits (« online knapsack problem »). Dans le cas à positions multiples, ils cherchent toujours maximiser les profits, mais en choisissant au plus une position par mot-clé (« online multiple-choice knapsack problem »). Ils proposent des stratégies permettant de déterminer la valeur d'enchère à fixer pour chaque mot-clé, puis ils fournissent des bornes de performance sur chacun de ces algorithmes.

Borgs et al. (2007) présentent une méthode simple et efficace qui tient compte des différences de rendement associées aux variations de la position d'une annonce. Leur modélisation considère que plusieurs annonceurs sont en compétition pour un même mot-clé et que chacun de ces annonceurs a un budget limité. L'heuristique proposée consiste à ajuster, de façon itérative, les enchères des mots-clés de façon à obtenir des retours sur investissement (RSI) marginaux égaux sur l'ensemble des mots-clés de la campagne.

Cette méthode exploite la propriété que le retour sur investissement (profit/coût) associé à un clic croît lorsque la position augmente. En effet, le profit associé à un clic est le même peu importe la position et le coût par clic décroît lorsque la position augmente. Cependant, le nombre de clics obtenu diminue lorsque la position croît. Ainsi, il faut trouver un compromis entre un volume de clics suffisamment élevé et un retour sur investissement satisfaisant.

Un annonceur qui désire maximiser le retour sur investissement global de sa campagne devra donc faire en sorte que le RSI marginal (i.e. dérivée de la fonction de revenu par rapport au coût) de chacun de ses mots-clés soit le même. Le cas contraire impliquerait qu'il existe au moins un mot-clé à RSI marginal supérieur à la moyenne et un autre à RSI inférieur à la moyenne. Dans ce cas, l'annonceur aurait avantage à diminuer son enchère pour le mot-clé à faible RSI marginal et

augmenter son enchère pour celui à fort RSI marginal. Bref, un équilibre est atteint lorsque tous les RSI marginaux sont égaux. Puisque le RSI marginal d'un mot-clé est difficile à estimer, Borgs et al. (2007) l'approximent avec le RSI, soit le ratio du revenu généré par le mot-clé divisé par son coût total. Cette approximation équivaut à utiliser un taux constant plutôt que d'effectuer un calcul de dérivée.

Ils utilisent également un mécanisme d'ajustement pour tenir compte de la contrainte de budget. Les valeurs d'enchères sont réajustées quotidiennement en fonction de la portion du budget total qui a été dépensée; ils multiplient la valeur de l'enchère des mots-clés par un facteur $R_i(t)$ qui dépend des résultats de la journée précédente $t$. Si le budget n'a pas été complètement dépensé au cours de la journée, la valeur des enchères est augmentée en utilisant $R_i(t) > 1$. Si la totalité du budget a été consommée en moins d'une journée, il faut diminuer les valeurs d'enchères en utilisant $R_i(t) < 1$.

Globalement, cette méthode est intéressante à cause de sa simplicité. Puisqu'elle procède par tâtonnements, elle ne nécessite qu'une quantité minimale d'information. Contrairement à plusieurs autres méthodes, il n'est pas nécessaire de connaître le coût associé aux positions autres que celle occupée actuellement puisque les valeurs d'enchères sont ajustées en fonction des RSI, qui sont calculés à la fin de chacune des journées.

Parallèlement, Feldman, Muthukrishnan, Pal & Stein (2008) décrivent une stratégie selon laquelle les valeurs d'enchères sont uniformes sur tous les mots-clés, c'est-à-dire qu'ils fixent la même valeur d'enchère peu importe le mot-clé. Ils visent à maximiser le nombre de clics obtenus par les annonces et cherchent une valeur d'enchère unique telle que l'espérance des coûts totaux n'excède pas le budget quotidien accordé.

Pour trouver cette valeur unique de CPC à fixer, ils agrègent d'abord les données de tous les mots-clés en sommant leurs coûts et leurs nombres de clics. Puis, ils génèrent le profil d'enchères (appelé « bid landscape » dans l'article), qui est essentiellement un graphique des combinaisons *nombre de clics – coût total* possibles. Sur le graphique agrégé, la solution est donnée par la valeur de coût total correspondant au budget global de la campagne. La stratégie d'enchère résultante est obtenue en utilisant une combinaison linéaire des deux points qui bordent la valeur du budget global. À partir du coût total et du nombre de clics associé à chacun de ces deux points,

il est facile d'en déduire deux valeurs d'enchère CPC idéales et la proportion du temps pour laquelle chacune des deux valeurs doit être appliquée au cours d'une journée.

Finalement, Kitts & Leblanc (2004) présentent une méthode beaucoup plus sophistiquée que les précédentes, mais qui nécessite la possession d'un historique de données suffisant pour chacun des mots-clés. Ils utilisent des méthodes de régression statistique pour estimer le nombre de clics en fonction de la position, ainsi que la position en fonction de la valeur d'enchère. L'utilisation de ces fonctions leur permet ultimement d'utiliser un modèle global d'optimisation en nombres entiers qui maximise les profits sous une contrainte de budget, en utilisant des variables qui correspondent aux valeurs d'enchère potentielles de chaque mot-clé.

Ainsi, l'optimisation est précédée par une phase d'analyse statistique qui consiste à générer des prédictions pour chacun des mots-clés de la campagne. Afin d'accorder plus d'importance aux observations récentes, une pondération décroissante est appliquée aux données en fonction de leur ancienneté. De plus, les fonctions de prédiction doivent satisfaire certaines conditions de qualité avant d'être considérées comme utilisables dans le modèle. Les auteurs formulent donc certains critères d'évaluation des régressions, puis ils discutent de stratégies d'exploration des mots-clés permettant d'améliorer la qualité de celles-ci, dans les cas où les critères de qualité ne sont pas satisfaits.

Globalement, le modèle d'optimisation est assez simple et la difficulté réside essentiellement dans l'obtention de fonctions de prédiction fiables. Cette approche est particulièrement intéressante puisqu'elle met l'accent sur l'aspect statistique du problème, qui est souvent négligé dans les autres études. En effet, la qualité des solutions obtenues suite à l'optimisation dépend énormément de la précision des fonctions de prédictions utilisées.

# CHAPITRE 3 MODÈLE D'OPTIMISATION

Le principal reproche que nous pouvons faire aux méthodes publiées dans la littérature jusqu'à présent est que la majorité d'entre elles ne tiennent pas compte de l'ensemble des mots-clés d'un compte lorsqu'elles prennent des décisions quant aux valeurs d'enchère à fixer; il s'agit de méthodes qui optimisent localement plutôt que globalement. Puisque les algorithmes d'optimisation locale ne considèrent pas simultanément tous les mots-clés affectés par les contraintes du problème, ils fournissent presque toujours des résultats sous-optimaux. Particulièrement dans le cas de grandes campagnes publicitaires constituées de plusieurs milliers de mots-clés, les solutions fournies par les approches locales risquent de s'éloigner considérablement de l'optimum.

En constatant les gains potentiels associés à l'utilisation d'une méthode d'optimisation globale, nous avons choisi d'orienter notre recherche dans cette direction. Nous proposons donc le développement d'une méthode d'optimisation globale qui tiendra compte, lors de l'attribution des valeurs d'enchère, de tous les mots-clés influencés par les contraintes du problème.

## 3.1 Structure des campagnes publicitaires

Tel qu'expliqué à la section 1.9, la plateforme logicielle d'Acquisio est utilisée par plus de 300 agences publicitaires. Chacune de ces agences possède un ou plusieurs comptes; un compte différent est créé pour chacun des clients de l'agence. Chaque compte est constitué de plusieurs campagnes (« PPC Campaign »). C'est à ce niveau que les budgets quotidiens doivent être définis. En effet, pour chaque campagne, l'annonceur doit définir le montant maximal qu'il est prêt à dépenser au cours d'une journée. Une fois ce budget fixé, le moteur de recherche se chargera de diffuser les annonces associées aux mots-clés de cette campagne jusqu'à ce que le budget soit complètement écoulé. Le montant facturé à l'annonceur pour une campagne est donc toujours inférieur ou égal à son budget quotidien. Les campagnes sont également caractérisées par des options de ciblage du marché, autant au niveau géographique que linguistique. Par exemple, un annonceur qui choisit de cibler la population francophone du Canada verra ses annonces apparaître uniquement lors des requêtes effectuées sur des ordinateurs au Canada, avec des utilisateurs qui ont enregistré le français comme une de leurs langues de préférence.

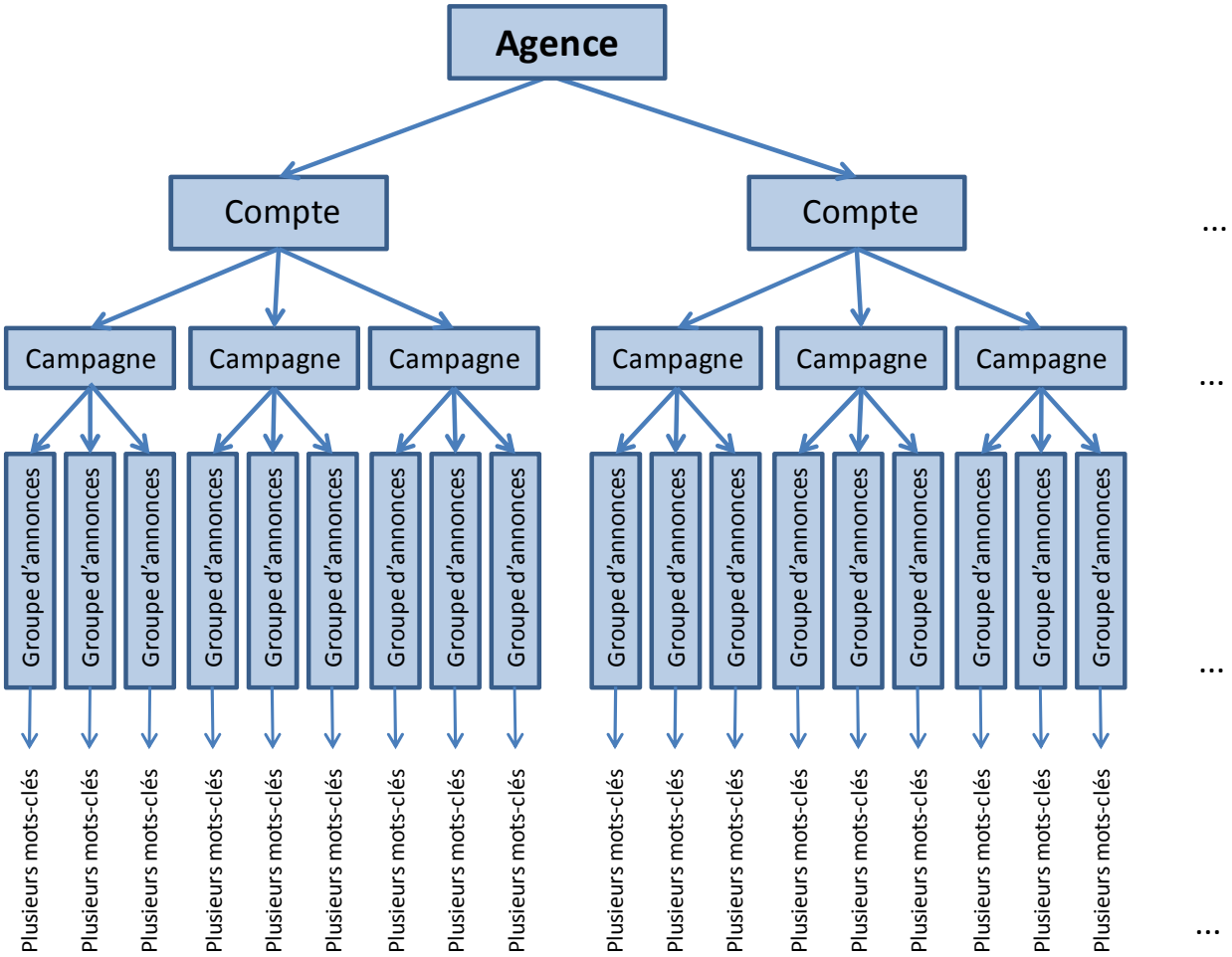



Figure 3.1 : Organisation des campagnes d'une agence publicitaire

Chacune des campagnes est constituée de plusieurs groupes d'annonces. Les groupes d'annonces fournissent la possibilité de cibler un marché en particulier, en permettant aux annonceurs d'associer différentes annonces textuelles à chacun de leurs mots-clés. En effet, chaque groupe d'annonces est constitué de plusieurs mots-clés et tous les mots-clés d'un même groupe d'annonces doivent utiliser les mêmes annonces textuelles. Ainsi, pour optimiser le rendement de leurs campagnes, les annonceurs ont intérêt à bien choisir les annonces textuelles qu'ils désirent associer à chacun de leurs mots-clés. Habituellement, les annonceurs associent environ 2 à 8 annonces textuelles différentes à chacun de leurs groupes d'annonces. Ils évaluent

périodiquement la performance de chacune de celles-ci et n'hésitent pas à en désactiver si leur rendement est considérablement plus faible que celui des autres annonces du groupe.

Ce sont les mots-clés qui constituent l'élément de base dans notre modèle d'optimisation. La stratégie que nous envisageons consisterait à choisir un ensemble de mots-clés à optimiser parmi tous les mots-clés d'un compte. De cette façon, l'annonceur aurait le choix d'optimiser son compte au complet ou exclure certains mots-clés s'il le désire. Ultimement, nous visons à automatiser l'attribution des valeurs d'enchère à chacun des mots-clés de façon à optimiser le rendement global de l'ensemble sélectionné.

## 3.2 Modèle initial

Pour résoudre ce type de problème, nous croyons qu'un modèle de programmation linéaire accompagné de méthodes de prédiction statistiques a le potentiel de fournir des solutions intéressantes. Si nous développons des méthodes capables de prédire le comportement de chacun des mots-clés avec des marges d'erreur relativement faibles, il est effectivement possible d'optimiser le rendement des campagnes publicitaires en utilisant une approche globale.

D'abord, il est important de clarifier ce que signifie « optimiser le rendement global ». Plus précisément, il faut déterminer l'objectif à maximiser ou à minimiser dans la fonction-objectif du programme linéaire. Tel que mentionné à la sous-section 1.6.1, la majorité des annonceurs qui gèrent des campagnes publicitaires sur les moteurs de recherche le font dans le but de générer des profits pour leur entreprise. Puisque les profits sont calculés en soustrayant les coûts aux revenus, l'approche d'optimisation idéale consisterait à maximiser la différence entre les revenus et les coûts d'une campagne publicitaire.

Pour compléter le modèle, il faut définir les contraintes du problème. Essentiellement, ce problème est limité par des contraintes de budget. On cherche à positionner chacun des mots-clés dans un des espaces publicitaires disponibles de façon à maximiser les revenus, tout en respectant une limite de budget quotidien. Pour arriver à positionner chacun des mots-clés dans les positions ciblées, il faudra avoir recours à des fonctions de prédiction qui permettront d'estimer la valeur d'enchère nécessaire pour l'atteinte de chacune des positions.

Afin de modéliser la situation, nous avons choisi de représenter le problème à l'aide de variables binaires de position. Chaque mot-clé possède autant de variables de position qu'il existe de positions potentielles. Selon la position à laquelle ce mot-clé sera affecté, une de ses variables prendra la valeur 1 et toutes les autres prendront la valeur 0. Il s'agit donc d'un programme linéaire de type « sac-à-dos binaire à choix multiple ». Ce type de problème est classé NP-difficile (Martello & Toth, 1990, p.77), donc dépendant de la taille, il pourrait être nécessaire d'avoir recours à des méthodes heuristiques pour le résoudre.

Évidemment, il est impossible de choisir avec exactitude la position à laquelle une annonce sera positionnée. Cependant, en utilisant des fonctions de prédiction de la valeur d'enchère en fonction de la position, nous pourrons estimer la valeur d'enchère nécessaire pour obtenir une position donnée. En fixant les valeurs d'enchère intelligemment, il sera possible de maximiser la probabilité de se retrouver dans les positions souhaitées.

Dans cette section, nous fournissons d'abord une description des variables, des paramètres et des fonctions utilisés pour modéliser le problème. Par la suite, le programme linéaire est présenté, accompagné d'une description et de quelques remarques.

Variables

$y_{mp}$ = $\begin{cases} \text{1 si le mot-clé } m \text{ est affecté à la position } p, \\ \text{0 sinon} \end{cases}$

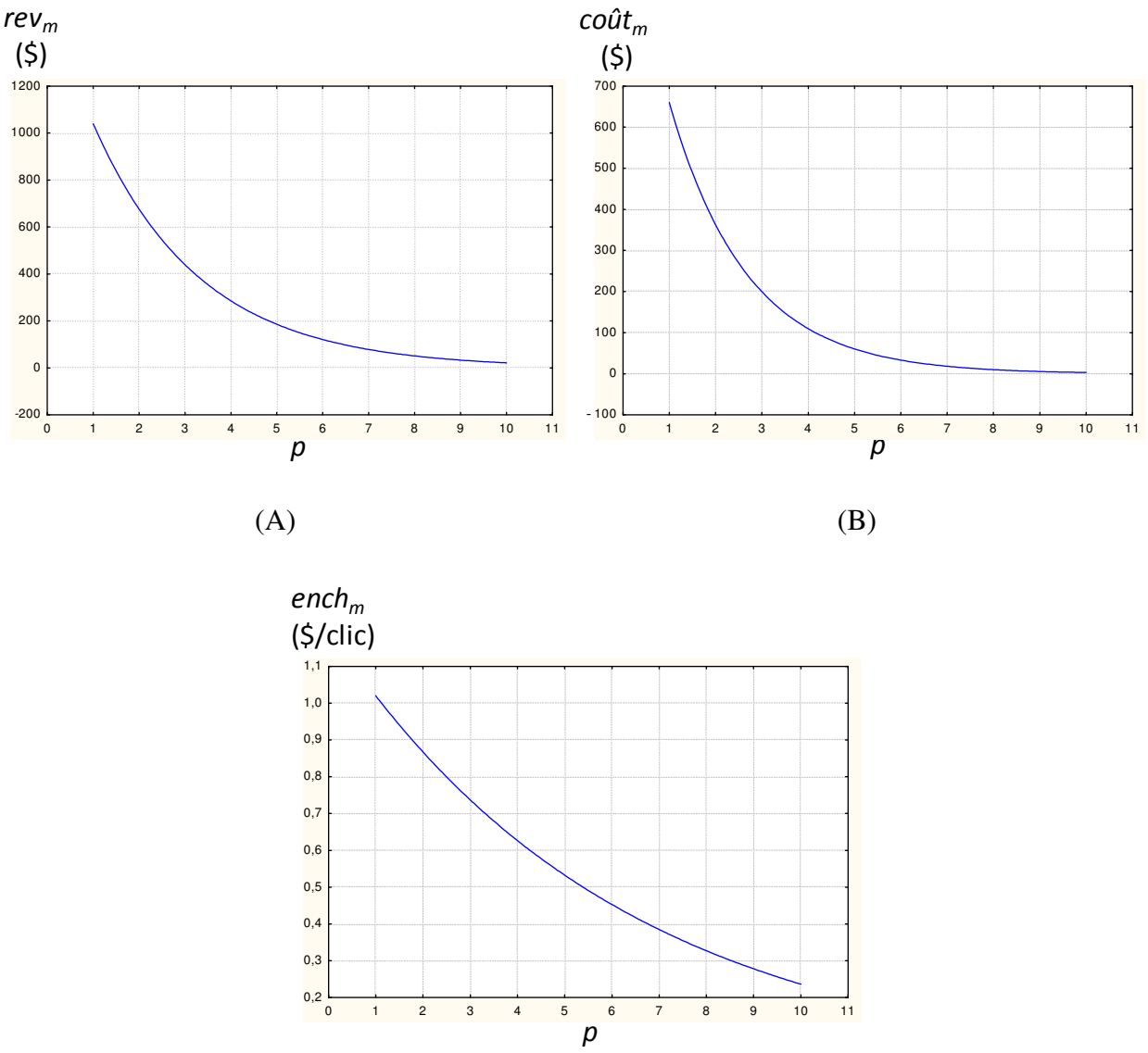
Paramètres et fonctions

| | |
|---|---|
| $B$ | Budget quotidien affecté à l'ensemble de mots-clés à optimiser |
| $rev_m(p)$ | Fonction de prédiction du revenu total (quotidien) obtenu par le mot-clé $m$ à la position $p$ |
| $coût_m(p)$ | Fonction de prédiction du coût total (quotidien) généré par le mot-clé $m$ à la position $p$ |
| $ench_m(p)$ | Fonction de prédiction de la valeur d'enchère nécessaire pour que l'annonce du mot-clé $m$ soit placée à la position $p$ |

La Figure 3.2 montre un exemple de fonctions de prédiction qui pourraient être obtenues pour un mot-clé donné. Dans cet exemple, nous utilisons des fonctions strictement décroissantes. La forme décroissante de la fonction $co\hat{u}t_m(p)$ s'explique par le fait que les clics et les CPC moyens sont généralement décroissants lorsque la position augmente. De plus, la forme décroissante de la fonction $ench_m(p)$ est justifiée par le fonctionnement de l'algorithme GSP qu'utilisent les moteurs de recherche pour classer les annonces; pour un même annonceur et un même indice de qualité, les valeurs d'enchère sont toujours décroissantes en fonction de la position. Pour ce qui est de la fonction $rev_m(p)$, sa forme exacte peut varier d'un mot-clé à l'autre; dépendant du contexte, il est possible que la fonction ne soit pas décroissante sur tout son domaine.

Plusieurs méthodes de prédiction différentes pourraient être utilisées pour prédire les revenus, coûts et valeurs d'enchère en fonction de la position. Les méthodes les plus populaires dans la littérature sont des méthodes qui impliquent des régressions. En appliquant de telles méthodes, les fonctions de prédiction obtenues sont des fonctions continues, avec lesquelles il est possible d'estimer les valeurs des variables même pour des valeurs de positions non entières. Cependant, pour la résolution de notre programme linéaire, il n'est pas nécessaire d'avoir des fonctions de prédiction continues. En effet, puisque les variables de position utilisées ne ciblent que les valeurs de position entières, il serait suffisant de discrétiser les fonctions et déclarer un certain nombre de paramètres pour chacun des mots-clés $m$.

Le nombre de paramètres à utiliser pour chaque mot-clé dépendra du nombre de positions $p$ que nous choisissons de considérer. Généralement, les références dans la littérature ne considèrent que 8 à 11 positions, car il s'agit du nombre d'annonces qui sont affichées sur la première page de résultats. Les volumes de recherche étant habituellement beaucoup plus faibles sur les pages subséquentes, il est très fréquent de les négliger dans les modèles. Dans notre cas, nous utiliserons une valeur de position maximale de 10. Bref, les valeurs de $p$ vont varier de 1 à 10, ce qui fait en sorte que chaque mot-clé $m$ aura 10 paramètres $rev_{mp}$, 10 paramètres $co\hat{u}t_{mp}$ et 10 paramètres $ench_{mp}$.




(A) (B)

(C)

Figure 3.2 : Exemples de fonctions de prédiction : revenus (A), coûts (B) et valeurs d'enchère (C) en fonction de la position moyenne

Discrétisation des fonctions

$rev_{mp}$ Revenu total (quotidien) prédit pour le mot-clé $m$ à la position $p$

$coût_{mp}$ Coût total (quotidien) prédit pour le mot-clé $m$ à la position $p$

$ench_{mp}$ Valeur d'enchère nécessaire pour que l'annonce du mot-clé $m$ soit placée à la position $p$

Par exemple, avec les fonctions montrées à la Figure 3.2, les valeurs de paramètres suivantes seraient obtenues :

Tableau 3.1 : Exemple de discrétisation de fonctions de prédiction (revenus, coûts et valeurs d'enchère en fonction de la position)

| $p$ | $rev_{mp}$ | $coût_{mp}$ | $ench_{mp}$ |
|---|---|---|---|
| 1 | 1040,00 | 660,00 | 1,02 |
| 2 | 676,00 | 363,00 | 0,87 |
| 3 | 439,40 | 199,65 | 0,74 |
| 4 | 285,61 | 109,81 | 0,63 |
| 5 | 185,65 | 60,39 | 0,53 |
| 6 | 120,67 | 33,22 | 0,45 |
| 7 | 78,44 | 18,27 | 0,38 |
| 8 | 50,98 | 10,05 | 0,33 |
| 9 | 33,14 | 5,53 | 0,28 |
| 10 | 21,54 | 3,04 | 0,24 |

Programme linéaire (quotidien)

*Fonction-objectif :*

$$max \sum_{m} \sum_{p} \left(rev_{mp} - co\hat{u}t_{mp}\right) * y_{mp}$$

*Contraintes :*

$$\sum_{p} y_{mp} = 1 \qquad \forall m$$

$$\sum_{m} \sum_{p} co\hat{u}t_{mp} * y_{mp} \leq B$$

$$y_{mp} \in \{0,1\} \qquad \forall m, \forall p$$

Utilisation de la solution

Une fois que le programme linéaire est résolu, chaque mot-clé considéré par le modèle est affecté à une position optimale. Il suffit alors d'utiliser les valeurs $ench_{mp}$ pour fixer les valeurs d'enchère et ainsi essayer d'atteindre les positions ciblées.

Description du modèle

La fonction-objectif vise à maximiser les profits, qui sont obtenus en soustrayant les coûts estimés aux revenus estimés. En fonction des valeurs qui seront attribuées aux variables de position de chaque mot-clé, il sera possible de calculer la somme des profits associée à une solution.

Le premier ensemble de contraintes vise à s'assurer que chaque mot-clé soit affecté à une seule position. Puisque les variables de position sont binaires, il suffit de contraindre la somme de la valeur des variables de position d'un mot-clé à 1.

La deuxième contrainte assure que la somme des coûts espérés n'excède pas le budget quotidien qui a été défini pour l'étendue des mots-clés à optimiser. En sommant les coûts de chaque mot-clé, il est possible d'estimer les coûts totaux pour une journée.

Finalement, le dernier ensemble de contraintes définit les variables de position comme étant binaires. Par conséquent, elles ne peuvent que prendre des valeurs de 0 ou 1.

## 3.3 Limites du modèle

Le modèle précédent est celui qui, en théorie, maximiserait les profits des campagnes publicitaires. À première vue, le modèle peut sembler relativement simple et facile à résoudre. Cependant, en analysant les données de plusieurs mots-clés provenant des diverses banques de données étudiées, nous avons soulevé quelques problèmes qui feraient en sorte que cette approche ne fonctionnerait pas en pratique.

D'abord, plusieurs annonceurs ne tiennent pas compte du revenu associé à chacune des conversions qu'ils obtiennent. Cela est souvent dû à la nature de leur entreprise; il n'est pas toujours possible de déterminer la valeur associée à une conversion. En effet, les conditions d'obtention d'une conversion sont définies par les annonceurs et elles ne sont pas uniquement associées à des achats. Par exemple, une conversion peut être obtenue suite à une inscription, un visionnement, une demande d'information, etc. Dans ces cas, il peut être difficile d'associer une valeur monétaire précise à chacune des conversions obtenues.

De plus, plusieurs annonceurs ne comptabilisent pas les valeurs de revenu associées à chacune des conversions qu'ils obtiennent. En effet, les annonceurs n'ont pas toujours des mécanismes en place pour stocker des données relatives à leurs conversions. Ainsi, ils n'effectuent aucun suivi sur la valeur monétaire associée à chacune de leurs conversions et il est donc impossible de modéliser le comportement du revenu en fonction de la position.

Ensuite, la rareté des conversions fait en sorte que le nombre d'observations n'atteint presque jamais des seuils significatifs pour obtenir des prédictions de qualité. Puisque les revenus sont associés aux conversions, des valeurs de revenu sont enregistrées uniquement lorsqu'une conversion est obtenue. Or, dans la plupart des campagnes publicitaires, les conversions n'ont lieu que très rarement. En effet, les taux de conversion moyens des campagnes publicitaires ont

généralement des valeurs de l'ordre de 0,1% à 1%[2], donc même les mots-clés qui génèrent plusieurs centaines de clics par jour n'obtiennent pas plus que quelques conversions.

Le Tableau 3.2 montre un exemple de 60 jours de données provenant d'un mot-clé à haut volume extrait de la campagne d'un des clients d'Acquisio. Supposons que nous voulions modéliser les revenus générés par ce mot-clé en fonction de la position. Puisque les données de revenu de cette entreprise ne sont pas comptabilisées, nous posons l'hypothèse que chaque conversion rapporte en moyenne la même valeur de revenu. La Figure 3.3 montre la répartition des conversions en fonction de la position moyenne pour ce mot-clé. Nous constatons que les valeurs de conversion non nulles ne sont pas suffisamment abondantes pour obtenir des fonctions de prédiction fiables. De plus, puisque les valeurs de conversions sont limitées à un ensemble de valeurs entières, nous observons un phénomène de formation de plateaux de points dans les graphiques lorsque le nombre moyen de conversions est faible.

Sachant que le mot-clé présenté dans cet exemple possède un volume de clics et de conversions relativement élevé comparativement à la moyenne, nous arrivons à la conclusion que le manque de conversions sera un problème pour la forte majorité des mots-clés d'une campagne. Bref, ce mot-clé démontre à quel point les conversions sont rares et pourquoi il est impossible de générer des fonctions de prédiction $rev_m(p)$ en utilisant des méthodes de régression.

Finalement, il n'est pas toujours souhaitable d'optimiser une campagne publicitaire uniquement en fonction des revenus générés par les mots-clés, car un tel modèle ne tient pas compte du cycle de consommation au complet. Dans plusieurs cas, le consommateur exécutera plusieurs requêtes vagues avant de converger vers une requête plus précise à laquelle sera attribuée la conversion (et par conséquent, les revenus). Par exemple, un client qui veut acheter un ordinateur pourra débuter ses recherches avec des requêtes comme « ordinateur », « ordinateurs à vendre » et « acheter ordinateur ». Une fois qu'il aura trouvé une marque qui lui plaît, il pourra alors se renseigner sur les différents modèles offerts par cette marque en exécutant des requêtes telles que « ordinateur de marque X », « ordinateurs de marque X à vendre » et « acheter ordinateur de marque X ». Une fois son choix effectué, il pourrait retourner effectuer une requête très précise comme « ordinateur modèle Y numéro de série Z ». Lorsqu'il effectuera son achat, la conversion ne sera

[2] Basé sur des valeurs calculées à partir des bases de données disponibles chez Acquisio

enregistrée que pour le dernier mot-clé recherché (« ordinateur modèle Y numéro de série Z »). Cela fait en sorte que les requêtes plus spécifiques seront favorisées par rapport aux requêtes plus vagues, même si ces dernières sont parfois essentielles au bon fonctionnement d'une campagne publicitaire.

Tableau 3.2 : Exemple de données historiques (position moyenne, nombre de clics et nombre de conversions) d'un mot-clé typique

| Jour | Position moyenne | Nombre de clics | Nombre de conversions | Jour | Position moyenne | Nombre de clics | Nombre de conversions |
|---|---|---|---|---|---|---|---|
| 1 | 4,28 | 115 | 2 | 31 | 6,69 | 48 | 0 |
| 2 | 4,16 | 58 | 2 | 32 | 6,43 | 31 | 1 |
| 3 | 4,58 | 50 | 0 | 33 | 5,36 | 105 | 0 |
| 4 | 4,26 | 84 | 2 | 34 | 5,82 | 92 | 2 |
| 5 | 4,80 | 81 | 0 | 35 | 5,80 | 94 | 0 |
| 6 | 3,40 | 106 | 5 | 36 | 7,37 | 54 | 0 |
| 7 | 5,40 | 100 | 0 | 37 | 7,29 | 24 | 0 |
| 8 | 4,88 | 47 | 1 | 38 | 8,63 | 14 | 0 |
| 9 | 6,92 | 15 | 0 | 39 | 11,00 | 3 | 1 |
| 10 | 6,96 | 18 | 0 | 40 | 6,30 | 59 | 1 |
| 11 | 6,21 | 43 | 0 | 41 | 5,48 | 56 | 1 |
| 12 | 7,12 | 30 | 0 | 42 | 5,28 | 78 | 1 |
| 13 | 6,87 | 39 | 0 | 43 | 5,02 | 84 | 0 |
| 14 | 6,77 | 40 | 0 | 44 | 4,98 | 39 | 3 |
| 15 | 6,63 | 55 | 0 | 45 | 5,39 | 75 | 1 |
| 16 | 7,92 | 33 | 1 | 46 | 6,24 | 75 | 1 |
| 17 | 6,90 | 60 | 0 | 47 | 6,19 | 43 | 0 |
| 18 | 6,02 | 63 | 0 | 48 | 5,99 | 52 | 1 |
| 19 | 6,52 | 50 | 0 | 49 | 5,81 | 58 | 1 |
| 20 | 5,99 | 68 | 2 | 50 | 5,94 | 81 | 0 |
| 21 | 4,99 | 76 | 1 | 51 | 6,46 | 71 | 1 |
| 22 | 5,25 | 128 | 1 | 52 | 7,73 | 18 | 3 |
| 23 | 5,53 | 74 | 0 | 53 | 6,23 | 44 | 0 |
| 24 | 5,35 | 55 | 1 | 54 | 7,88 | 25 | 0 |
| 25 | 5,68 | 70 | 0 | 55 | 7,43 | 25 | 0 |
| 26 | 6,07 | 54 | 0 | 56 | 7,67 | 21 | 0 |
| 27 | 6,75 | 14 | 1 | 57 | 6,82 | 21 | 0 |
| 28 | 7,32 | 7 | 0 | 58 | 6,73 | 18 | 0 |
| 29 | 9,12 | 5 | 0 | 59 | 6,48 | 27 | 1 |
| 30 | 10,66 | 11 | 1 | 60 | 3,33 | 136 | 2 |

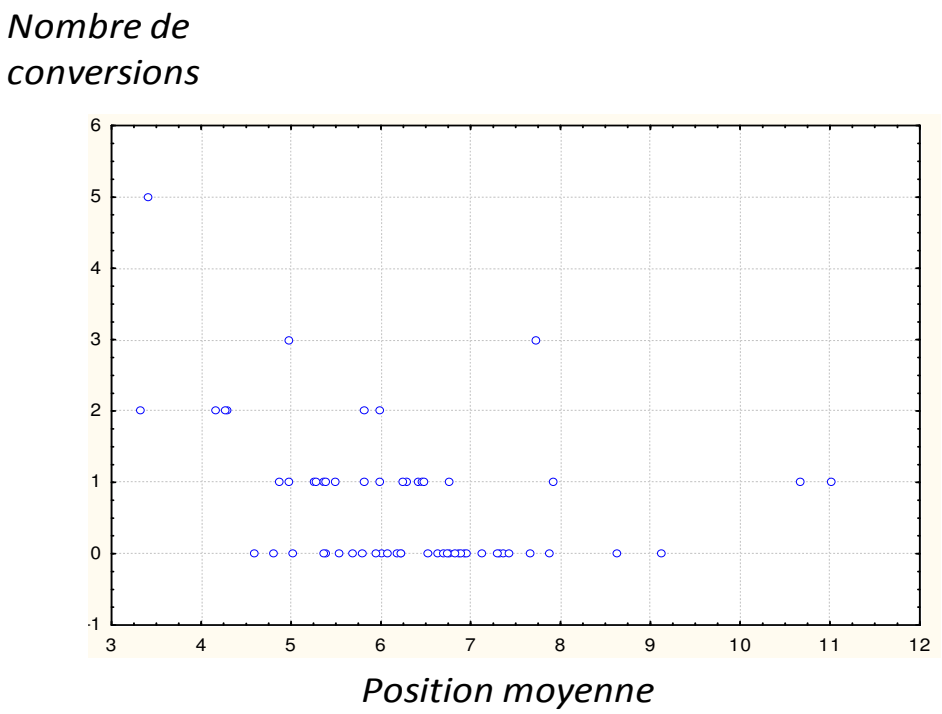



Figure 3.3 : Graphique du nombre de conversions en fonction de la position moyenne pour un mot-clé

Considérant tous ces problèmes associés à l'optimisation des campagnes publicitaires en fonction des profits, nous concluons qu'il serait préférable de changer notre objectif. En effet, il est très difficile d'obtenir des fonctions de prédiction $rev_{mp}$ fiables pour chacun des mots-clés d'une campagne publicitaire. De plus, il est difficile d'attribuer, à chacun des mots-clés, une valeur de revenu adéquate qui tient compte de son influence dans le cycle d'achat complet. Nous en concluons donc qu'il faudra modifier notre modèle d'optimisation.

## 3.4 Modèle alternatif

En considérant les limites mentionnées à la section précédente, nous arrivons à la conclusion qu'il est nécessaire de modifier le modèle d'optimisation pour obtenir un algorithme fonctionnel et efficace. Puisque l'impossibilité de modéliser adéquatement les revenus est la principale cause du problème, il faut d'abord se définir une nouvelle fonction-objectif. Nous simplifions donc le

problème en choisissant de maximiser le nombre de clics plutôt que les profits. Il est intéressant pour un annonceur de maximiser le nombre de clics qu'il obtient, car il maximise alors ses opportunités d'offrir ses produits ou services à sa clientèle cible. Ainsi, nous pouvons donc considérer que le nombre de clics obtenu à budget constant est un indicateur indirect du niveau de profitabilité d'une campagne.

D'un point de vue d'analyse statistique, il est plus intéressant de considérer les clics que les profits, car les fonctions du nombre de clics selon la position seront plus faciles à obtenir que les fonctions du revenu selon la position. Les clics sont beaucoup plus abondants que les conversions (environ 100 à 1000 fois plus abondants, dépendant des taux de conversion moyens), donc il devrait être possible de leur appliquer des méthodes de prédiction telles que les régressions, dans la majorité des cas. Bien-sûr, il existera toujours certains mots-clés avec un nombre de clics plus limité, mais ils sont beaucoup plus rares que les mots-clés qui ont un nombre de conversions limité. Bref, puisque leur occurrence est beaucoup moins sporadique que celle des conversions, nous croyons qu'il sera plus facile de prédire le comportement des clics avec les méthodes de prédiction que nous envisageons.

La Figure 3.4 illustre la répartition du nombre de clics en fonction de la position pour le mot-clé qui a été présenté au Tableau 3.2. Afin de montrer la tendance décroissante, les observations ont été accompagnées d'une courbe de régression exponentielle. Dans la Figure 3.3, il était impossible de détecter une tendance dans les données historiques de conversion pour ce mot-clé, à cause de leur faible quantité. Cependant, le potentiel de prédiction est bien meilleur lorsque nous modélisons le nombre de clics.

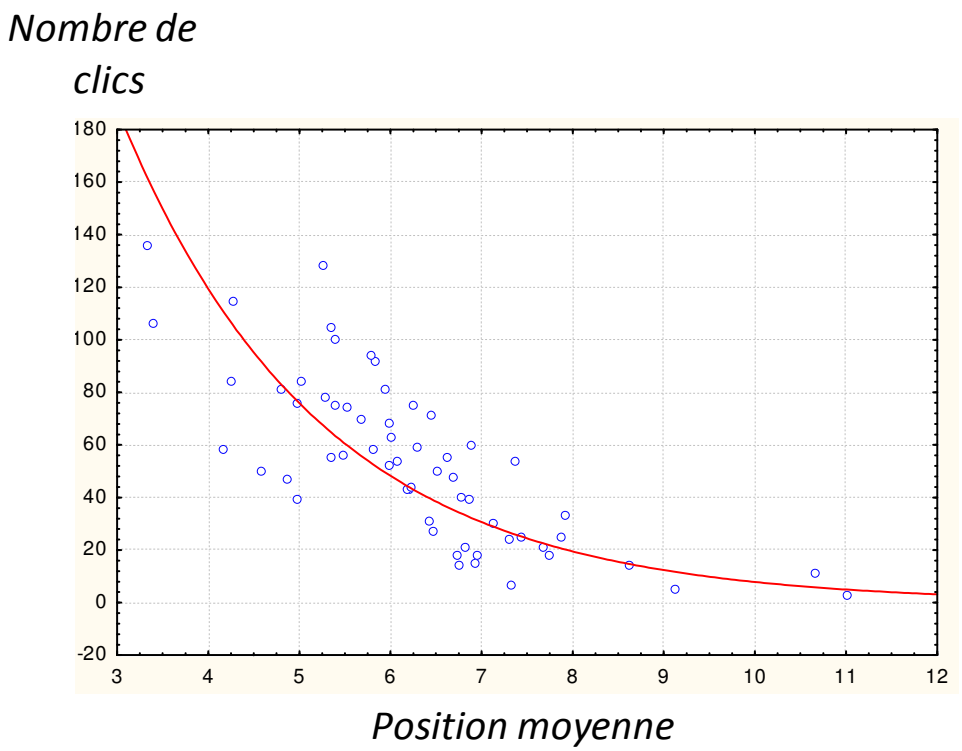



Figure 3.4 : Graphique du nombre de clics en fonction de la position moyenne pour un mot-clé

En somme, dans le programme linéaire d'optimisation, nous avons choisi le nombre total de clics comme fonction à maximiser. Pour appuyer cette décision, nous croyons important de mentionner que plusieurs références publiées dans la littérature choisissent également cet objectif. Notamment, Kitts et al. (2005), Feldman et al. (2008), Feldman & Muthukrishnan (2008) ainsi que Muthukrishnan et al. (2010) présentent des approches qui visent à maximiser les clics.

Le modèle que nous proposons est donc une variante de celui qui a été présenté à la section 3.2. Essentiellement, nous remplaçons la fonction de profits (revenus totaux – coûts totaux) par une fonction de clics et nous utilisons une fonction de CPC moyen plutôt qu'une fonction de coût total. Le nouveau modèle proposé est le suivant :

Variables

$y_{mp}$ = $\begin{cases} \text{1 si le mot-clé } m \text{ est affecté à la position } p, \\ \text{0 sinon} \end{cases}$

Paramètres et fonctions

| | |
|---|---|
| $B$ | Budget quotidien affecté à l'ensemble de mots-clés à optimiser |
| $clics_m(p)$ | Fonction de prédiction du nombre total de clics (quotidien) obtenus par le mot-clé $m$ à la position $p$ |
| $cpc_m(p)$ | Fonction de prédiction du coût par clic moyen du mot-clé $m$ à la position $p$ |
| $ench_m(p)$ | Fonction de prédiction de la valeur d'enchère nécessaire pour que l'annonce du mot-clé $m$ soit placée à la position $p$ |

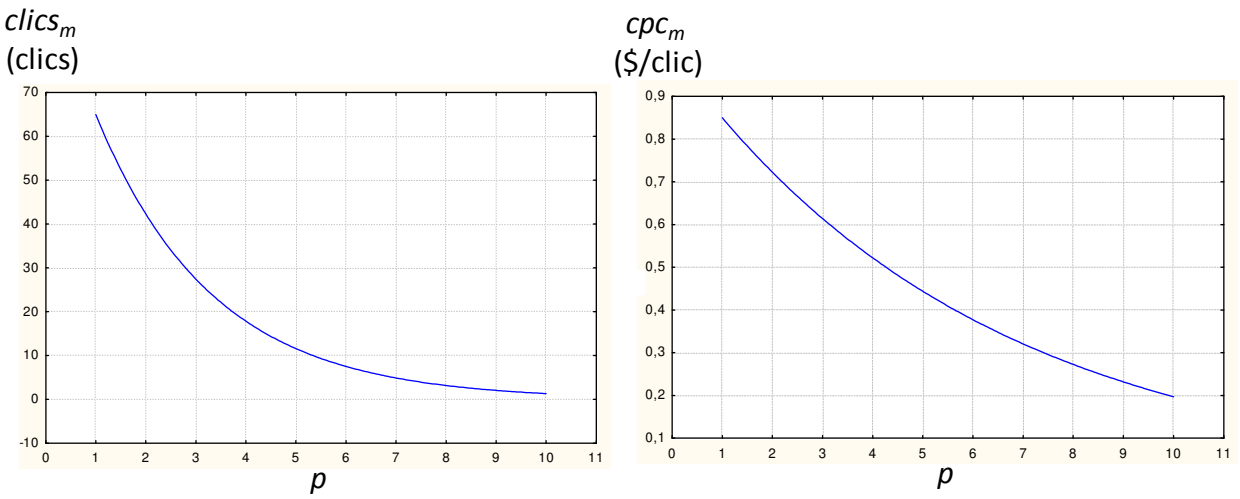
La Figure 3.5 fournit un exemple de fonctions de prédiction qui pourraient être obtenues pour un mot-clé donné.

Tout comme dans le modèle initial, nous discrétisons les valeurs de ces fonctions de prédiction pour obtenir des paramètres $clics_{mp}$, $cpc_{mp}$ et $ench_{mp}$.

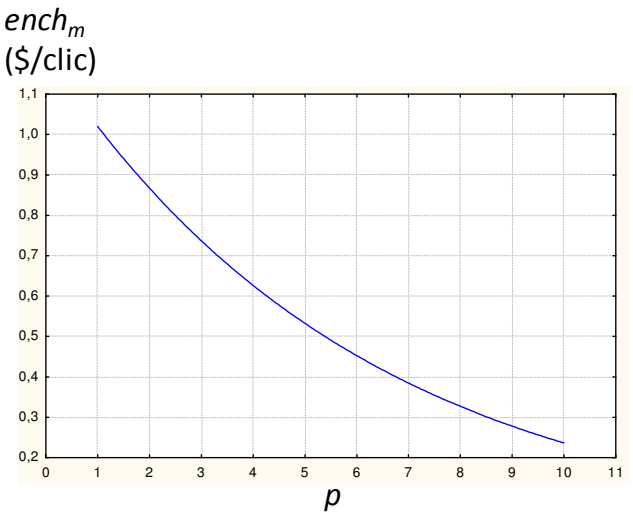
Discrétisation des fonctions

| | |
|---|---|
| $clics_{mp}$ | Nombre total de clics (quotidien) prédit pour le mot-clé $m$ à la position $p$ |
| $cpc_{mp}$ | Coût par clic moyen pour le mot-clé $m$ à la position $p$ |
| $ench_{mp}$ | Valeur d'enchère nécessaire pour que l'annonce du mot-clé $m$ soit placée à la position $p$ |

Par exemple, le Tableau 3.3 montre les valeurs de paramètres qui seraient obtenues en discrétisant les fonctions montrées à la Figure 3.5.




(A) (B)




(C)

Figure 3.5 : Exemples de fonctions de prédiction : nombre de clics (A), CPC moyens (B) et valeurs d'enchère (C) en fonction de la position moyenne

Tableau 3.3 : Exemple de discrétisation de fonctions de prédiction (nombre de clics, CPC moyens et valeurs d'enchère en fonction de la position)

| $p$ | $clics_{mp}$ | $cpc_{mp}$ | $ench_{mp}$ |
|---|---|---|---|
| 1 | 65,00 | 0,85 | 1,02 |
| 2 | 42,25 | 0,72 | 0,87 |
| 3 | 27,46 | 0,61 | 0,74 |
| 4 | 17,85 | 0,52 | 0,63 |
| 5 | 11,60 | 0,44 | 0,53 |
| 6 | 7,54 | 0,38 | 0,45 |
| 7 | 4,90 | 0,32 | 0,38 |
| 8 | 3,19 | 0,27 | 0,33 |
| 9 | 2,07 | 0,23 | 0,28 |
| 10 | 1,35 | 0,20 | 0,24 |

Programme linéaire (quotidien)

*Fonction-objectif :*

$$max \sum_m \sum_p clics_{mp} * y_{mp}$$

*Contraintes :*

$$\sum_p y_{mp} = 1 \qquad \forall m$$

$$\sum_m \sum_p clics_{mp} * cpc_{mp} * y_{mp} \leq B$$

$$y_{mp} \in \{0,1\} \qquad \forall m, \forall p$$

Utilisation de la solution

Une fois que le programme linéaire est résolu, chaque mot-clé considéré par le modèle est affecté à une position optimale. Tout comme dans le modèle initial, il suffit d'utiliser les valeurs $ench_{mp}$ pour fixer les valeurs d'enchère et ainsi tenter d'atteindre les positions ciblées.

Description du modèle

Tel qu'expliqué précédemment, la fonction-objectif vise simplement à maximiser le nombre de clics espéré. En utilisant les fonctions de prédiction du nombre de clics, il sera possible de calculer le nombre total de clics espéré en fonction des valeurs qui auront été attribuées aux variables de position.

Le premier ensemble de contraintes vise à s'assurer que chaque mot-clé soit affecté à une seule position. Puisque les variables de position sont binaires, il suffit de contraindre la somme de la valeur des variables de position d'un mot-clé à 1.

La deuxième contrainte assure que la somme des coûts espérés n'excède pas le budget quotidien qui a été défini pour l'étendue des mots-clés à optimiser. Il suffit de multiplier le nombre de clics estimé par le coût moyen par clic estimé pour obtenir le coût quotidien espéré d'un mot-clé en une position donnée. En sommant les coûts de chaque mot-clé, il est possible d'estimer les coûts totaux pour une journée.

Finalement, le dernier ensemble de contraintes définit les variables de position comme étant binaires. Par conséquent, elles ne peuvent que prendre des valeurs de 0 ou 1.

Prédiction des valeurs d'enchère

Une fois que le modèle aura déterminé la position à laquelle il faudrait placer chaque mot-clé, il suffira d'utiliser $ench_{mp}$ pour déterminer la valeur d'enchère à fixer. Il est important de faire la distinction entre la fonction de valeur d'enchère et celle de coût par clic moyen; la valeur d'enchère est toujours supérieure ou égale au coût par clic moyen. Ainsi, il faudrait utiliser $cpc_{mp}$ pour estimer les coûts totaux dans le modèle (voir deuxième contrainte), mais $ench_{mp}$ pour estimer l'enchère à fixer suite à l'obtention d'une solution, afin d'atteindre les positions voulues.

Cependant, les données de valeurs d'enchère n'étant pas disponibles au moment de la rédaction de ce mémoire, nous avons été contraints à utiliser les $cpc_{mp}$ pour estimer les valeurs de $ench_{mp}$. Cela génère évidemment des erreurs de prédiction dans notre modèle, mais puisque la différence entre la valeur d'enchère et le coût par clic est généralement assez faible, nous considérons qu'il s'agit d'une approximation acceptable. C'est donc $cpc_{mp}$ qui sera utilisé pour orienter le choix des valeurs d'enchères lors du positionnement des annonces. Afin de réduire l'écart entre les

deux fonctions, il serait aussi possible de multiplier les $cpc_{mp}$ par un coefficient légèrement supérieur à 1 afin d'obtenir une estimation de $ench_{mp}$. Par exemple, si la valeur d'enchère est en moyenne 20% plus élevée que le CPC moyen, nous pourrions utiliser un coefficient multiplicatif de 1,20. Nous recommandons toutefois d'utiliser $ench_{mp}$ dans les cas où les données relatives aux valeurs d'enchère sont disponibles.

Choix de modélisation

Au niveau des choix de modélisation, nous avons opté pour un modèle à variables de position binaires. Cette option semble facile à appliquer, mais il aurait également été possible d'utiliser un modèle avec des variables de position continues. En effet, nous savons que les données de position moyenne sont des valeurs non entières, donc il n'est pas nécessaire de considérer un ensemble de positions discret lors de la modélisation. Avec des variables de position continues, il serait possible de cibler des positions non entières. Cependant, nous ne croyons pas que cela présente des avantages considérables, car l'estimation des valeurs d'enchère en fonction de la position n'est pas suffisamment précise pour justifier l'utilisation d'un modèle à variables continues. De plus, l'utilisation de variables continues pourrait rendre le modèle non linéaire, dans le cas où des fonctions exponentielles sont utilisées pour effectuer les prédictions. Cela risquerait de complexifier la résolution du programme.

Désactivation de mots-clés

Puisqu'il peut parfois être souhaitable de désactiver certains mots-clés (i.e. leur fixer une valeur d'enchère nulle), nous avons choisi d'introduire une variable de position supplémentaire $y_{m,11}$ pour chacun des mots-clés $m$. Ainsi, dans le cas de mots-clés peu performants, le modèle ne sera pas contraint à fixer des valeurs d'enchère nécessairement positives. Pour désactiver un mot-clé $m$, il suffira de fixer $y_{m,11} = 1$. De cette façon, la contrainte 1 fera en sorte que le mot-clé ne pourra pas être affecté à une des 10 positions considérées.

Contraintes additionnelles

Avec ce modèle, il serait très facile d'ajouter des contraintes additionnelles pour répondre aux besoins spécifiques de certains clients. Par exemple, dans le cas d'un client qui désire voir toutes ses annonces situées dans des positions de valeurs inférieures à 5, il serait possible de fixer $y_{mp} = 0$ pour toutes les positions $p \geq 5$. Par ailleurs, dans le cas d'un client qui veut payer au plus 1,00$ par clic, nous pourrions imposer la contrainte suivante : $cpc_{mp} * y_{mp} \leq 1{,}00 \ \ \forall m, \forall p$. Bref, le modèle est très flexible et il est possible de le modifier pour s'adapter à des situations particulières.

Fréquence de résolution

Le programme linéaire sera résolu périodiquement, afin de repositionner les annonces de façon optimale. La fréquence de résolution sera choisie stratégiquement par le gestionnaire de campagne. Il n'est donc pas nécessaire d'y inclure des indices de temps pour les variables et les paramètres. En effet, chaque valeur qui est donnée par les variables, paramètres et fonctions est une valeur quotidienne. Les fonctions de prédiction prédiront le nombre de clics et le coût par clic moyen d'une journée, pour une position donnée. Cela est cohérent avec les données historiques qui sont à notre disposition, car les moteurs de recherche fournissent les valeurs de nombre total de clics, coût par clic moyen et position moyenne sur une base quotidienne.

Étendue du modèle (« Scope »)

L'étendue de l'application du modèle est variable. Il est donc possible de l'appliquer à de petits groupes de mots-clés, tout comme il est possible de l'appliquer à une campagne publicitaire complète. Pour utiliser le modèle d'optimisation, il suffit de déterminer un sous-ensemble de mots-clés à optimiser et un budget associé à ce sous-ensemble. Bien-sûr, il est idéal d'avoir une étendue plus large, car une étendue étroite n'exploite pas totalement les bénéfices d'une optimisation globale. Par exemple, optimiser séparément 10 groupes de 100 mots-clés risque de fournir des solutions moins profitables qu'optimiser un seul groupe de 1000 mots-clés.

Importance de la qualité des fonctions de prédiction

Comme il est facile de constater, le modèle d'optimisation est relativement peu complexe. La fonction-objectif vise simplement à maximiser les clics et il suffit de respecter une contrainte de budget et une contrainte de position unique pour obtenir une solution réalisable. Ainsi, si nous sommes capables de prédire les coûts par clic moyens et le nombre de clics qui seront obtenus pour chaque mot-clé en chacune des positions potentielles, il devrait être possible de déterminer une allocation optimale du budget entre chacun des mots-clés.

Cependant, l'obtention de prédictions pour les fonctions $clics_m(p)$, $cpc_m(p)$ et $ench_m(p)$ de chacun des mots-clés à optimiser risque d'être une tâche difficile. En effet, puisque les données comportent généralement beaucoup de bruit, les méthodes de régression traditionnelles ne fournissent pas toujours des prédictions fiables. Afin de minimiser les erreurs de prédiction, il sera important de développer des méthodes de prédiction robustes et efficaces qui exploiteront toute l'information que nous fournissent les données historiques à notre disposition. Les chapitres qui suivent expliquent les différentes méthodes qui ont été testées dans le but d'obtenir des fonctions de prédiction de qualité.

# CHAPITRE 4 CLASSIFICATION DES MOTS-CLÉS

Tel qu'expliqué au chapitre précédent, nous souhaitons utiliser les fonctions $clics_m(p)$ et $cpc_m(p)$ pour prédire le comportement de chaque mot-clé en fonction de la position de son annonce. Théoriquement, le programme linéaire permettrait de modéliser l'ensemble des campagnes publicitaires et fournirait les valeurs d'enchère optimales pour chacun des mots-clés qui constituent celles-ci. Cependant, en pratique, la qualité des solutions fournies par le modèle dépend énormément de la qualité des prédictions pouvant être obtenues.

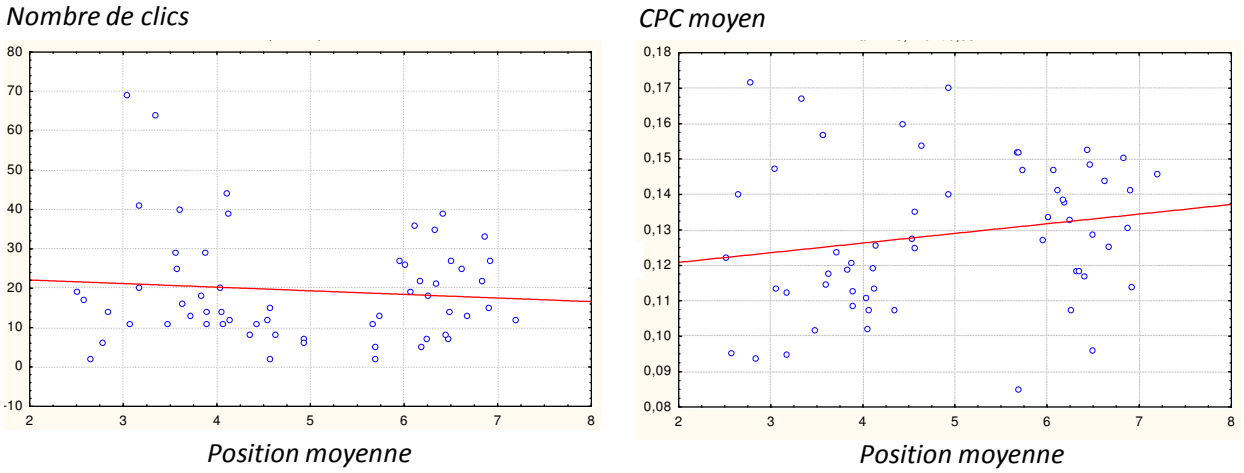
Initialement, nous avions l'intention d'appliquer des méthodes de régression aux données historiques des mots-clés dans le but de générer des fonctions de prédiction. Cependant, comme nous avons pu le constater, ces méthodes ne fournissent pas toujours des fonctions de prédiction fiables. Nous avons rencontré, dans plusieurs cas, des obstacles qui font en sorte qu'il n'est pas toujours possible d'utiliser des régressions pour prédire le comportement des variables de clics et CPC moyen. La section 4.1 présente en détail les principaux problèmes rencontrés et explique pourquoi ils affectent le potentiel de prédiction d'un mot-clé. Par la suite, la section 4.2 explique la structure de classification qui sera utilisée pour tenir compte de ces problèmes lors de l'affectation des mots-clés aux diverses méthodes de prédiction disponibles. Finalement, à la section 4.3, nous apportons des précisions concernant une des méthodes envisagées pour le traitement de certains mots-clés particuliers.

## 4.1 Problèmes rencontrés

Forte dispersion dans les données

La dispersion dans les données historiques des mots-clés fait en sorte qu'il est parfois presque impossible de détecter une tendance au niveau de la variation des clics et CPC moyens en fonction de la position. Un exemple d'un tel mot-clé est illustré à la Figure 4.1. En traçant ses graphiques de prédiction, nous constatons que les données sont tellement dispersées qu'il est impossible de trouver un ajustement de fonction (linéaire, exponentielle, logarithmique, etc.) qui modélise bien l'ensemble des observations. Dans de tels cas, il n'est pas souhaitable d'utiliser la fonction donnée par l'équation de régression à des fins de prédiction, car la dispersion des

données est beaucoup trop grande et les coefficients de détermination ont généralement des valeurs près de 0.




(A) (B)

Figure 4.1 : Exemples de graphiques de nombre de clics (A) et de CPC moyen (B) avec forte dispersion dans les données

Plusieurs facteurs peuvent contribuer à expliquer une telle dispersion dans les données. D'abord, à cause de la nature de certaines campagnes publicitaires et du type de produit ou service offert, certains mots-clés ont des comportements périodiques (périodicités hebdomadaires, mensuelles, saisonnières...). Dans ces cas, les données peuvent varier énormément d'une journée à l'autre. Plus précisément, la périodicité des campagnes publicitaires peut engendrer de grandes variations au niveau des volumes de requêtes, ce qui peut faire en sorte que deux journées à la même position peuvent générer un nombre de clics considérablement différent. De plus, si la compétition entre les annonceurs s'intensifie pendant ces périodes, il est également possible de constater de grandes variations au niveau des CPC moyens par position. Bref, le comportement périodique de certains mots-clés peut être une cause de la grande dispersion dans les données. Étant donné la taille de la plupart des campagnes publicitaires, il peut être très difficile de tenir compte de la périodicité de chaque mot-clé dans un modèle de prédiction.

Par ailleurs, l'agrégation des données sur une base quotidienne peut également être une source de bruit dans les données historiques. Puisque le CPC moyen est calculé en effectuant la moyenne de chacun des CPC payés pendant la journée et que la position d'une annonce peut changer d'une requête à l'autre, il n'est pas possible de connaître le montant exact nécessaire pour atteindre une position précise. Par exemple, si un mot-clé qui a obtenu 2 clics pendant une journée enregistre une position moyenne de 3,5 et un CPC moyen de 1,00$, il est impossible de connaître les deux positions réellement occupées et les deux CPC réellement facturés. Nous pourrions supposer qu'il a obtenu un clic en position 3 avec un CPC légèrement supérieur à 1,00$ et un autre clic en position 4 avec un CPC légèrement inférieur à 1,00$. Cependant, il est également possible que l'annonce n'ait jamais visité les positions 3 et 4 et que les CPC payés ne soient pas du tout près de 1,00$. Des clics aux positions 2 et 5 auraient fourni la même valeur de position moyenne et des CPC de 1,75$ et 0,25$ auraient donné le même CPC moyen. Ainsi, nous pouvons considérer que l'agrégation quotidienne des données risque de générer un certain manque de précision lors de nos tentatives de prédiction.

Finalement, l'effet humain a un impact non négligeable sur la qualité des prédictions qui peuvent être obtenues. En tentant de modéliser le nombre de clics obtenu par une annonce, nous visons, d'une certaine façon, à prédire le comportement humain. Évidemment, les êtres humains n'ont pas toujours un comportement facilement prévisible; une simple variation dans les conditions externes peut affecter leurs priorités de consommation. Par conséquent, la variation de facteurs externes relatifs au contexte économique, au contexte sociopolitique ou même à la météorologie peuvent influencer considérablement le rendement d'une campagne publicitaire. Puisqu'il est presque impossible de tenir compte de tous ces facteurs dans un modèle de prédiction, nous sommes contraints à accepter l'imprécision associée aux facteurs humains et son impact sur la qualité de nos fonctions de prédiction.

Nombre d'observations insuffisant

Une trop faible quantité de données historiques peut parfois faire en sorte qu'il soit impossible d'obtenir des fonctions de prédiction. En effet, il est fréquent de retrouver de nouveaux mots-clés qui possèdent un nombre très restreint d'observations. Dans ces cas, il n'est pas souhaitable

d'appliquer des méthodes de régression sur un nombre d'observations non significatif, car nous risquons de générer des prédictions à marges d'erreur très élevées.

Nombres moyens de clics trop faibles

Lorsqu'un graphique du nombre de clics en fonction de la position moyenne est tracé, la variable de position est une variable continue puisqu'elle peut prendre n'importe quelle valeur, entière ou non, dans l'intervalle de positions. Cependant, la variable du nombre de clics est une variable discrète, car les clics sont toujours fournis sous forme de valeurs entières. Ainsi, lorsque le nombre de clics obtenu est trop faible, les points du graphique ont tendance à se regrouper sous forme de plateaux. Dans ces cas, il est souvent difficile d'obtenir de bonnes régressions à partir des données.

La Figure 4.2 montre un tel exemple de graphique de clics en fonction de la position moyenne. Le problème observé est sensiblement le même que celui rencontré lors de la prédiction des conversions, décrit à la section 3.3. Malgré un nombre significatif d'observations, il est simplement impossible de trouver une régression qui s'ajuste bien à l'ensemble des points. Ce phénomène se produit assez fréquemment, car une campagne publicitaire est typiquement constituée d'une majorité de mots-clés à faible volume.

Plage de données de position trop restreinte

Lorsque les données de position moyenne (variable indépendante dans les régressions) ne couvrent pas un intervalle de positions suffisamment grand, il est souvent difficile d'observer une tendance dans les données. Le nombre de clics et les CPC moyens sont généralement décroissants lorsque la position moyenne augmente, mais il peut être difficile d'observer cette décroissance lorsque l'étendue de positions considérées n'est pas assez large. La Figure 4.3 illustre, en exemple, les données d'un mot-clé pour lequel ce problème est présent. Dans les deux graphiques, la position moyenne maximale considérée est 4,34 et la position moyenne minimale est 2,27; les données sont donc réparties sur une plage d'à peine deux positions. Évidemment, cette plage est beaucoup trop limitée pour qu'une régression puisse capturer l'effet de décroissance des clics et des CPC moyens.

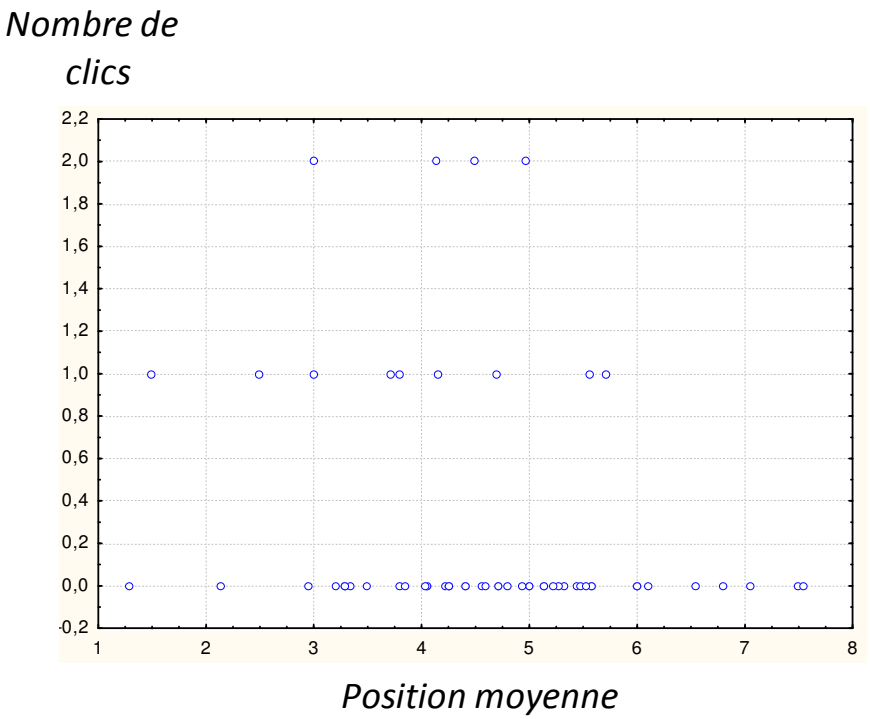
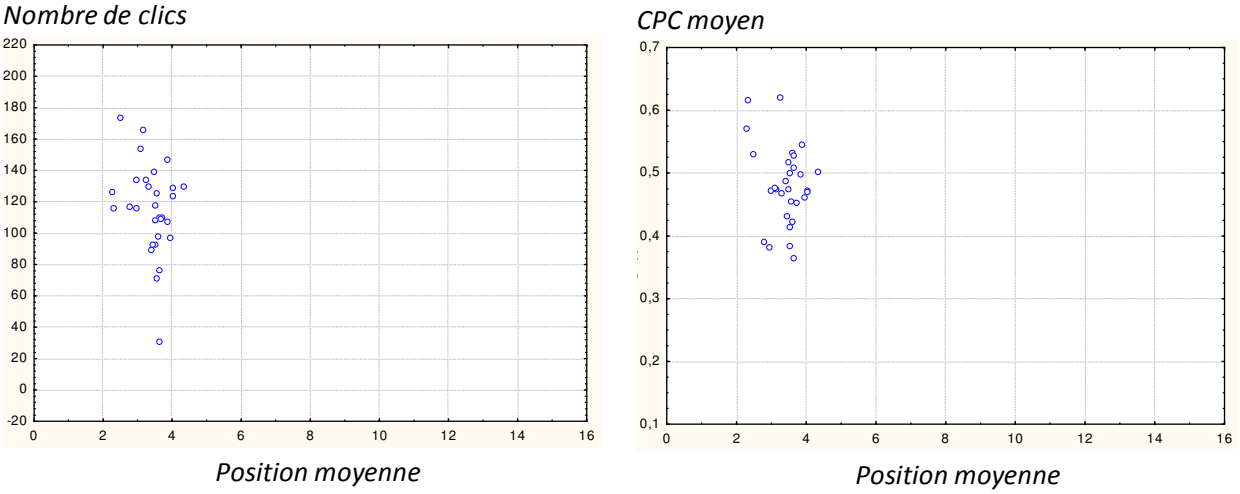


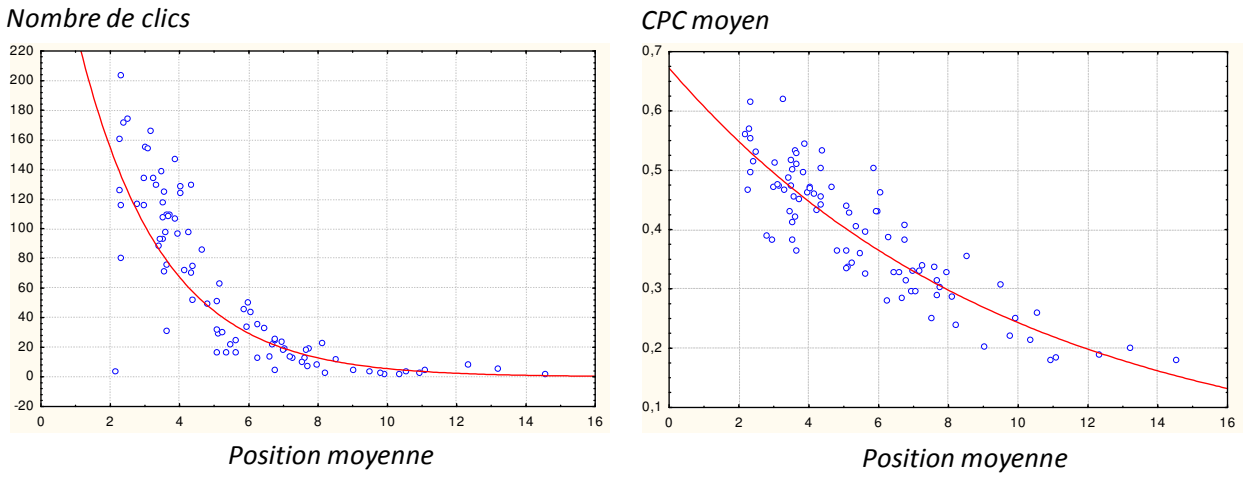

Figure 4.2 : Exemple de graphique de clics en fonction de la position moyenne pour un mot-clé à faible volume




(A) (B)

Figure 4.3 : Exemples de graphiques de nombre de clics (A) et de CPC moyen (B) en fonction de la position moyenne, avec plage de données de position trop restreinte

Lorsque nous considérons une étendue de positions plus grande dans l'historique de données, nous constatons que les tendances décroissantes des clics et CPC moyens en fonction de la position moyenne deviennent de plus en plus faciles à détecter. En prenant le mot-clé utilisé à l'exemple de la Figure 4.3, il est possible de montrer que l'ajout d'observations dans des positions plus variées permet d'obtenir de meilleures fonctions de prédiction. La Figure 4.4 montre ce qui est obtenu lorsque nous ajoutons 60 jours supplémentaires aux 30 jours de données qui ont été présentés à la Figure 4.3. Avec cet ajout, il est possible de couvrir une étendue de plus de 12 positions différentes, ce qui a pour effet de montrer clairement le rythme de décroissance du nombre de clics et des CPC moyens en fonction de la position moyenne.




(A) (B)

Figure 4.4 : Exemples de graphiques de nombre de clics (A) et de CPC moyen (B) en fonction de la position moyenne, avec plage de données de position suffisamment grande

À cause des méthodes de gestion d'enchères qui sont habituellement utilisées par les annonceurs, le phénomène d'étendue de positions très restreinte se produit très fréquemment. En effet, une stratégie communément utilisée dans le domaine consiste à utiliser des logiciels de gestion à base de règles afin de cibler des positions particulières. Plus spécifiquement, les annonceurs choisissent, d'une façon quelconque, la position à laquelle ils désirent retrouver leur annonce. Par la suite, en fonction de la position moyenne obtenue pour la journée, le logiciel augmente ou

diminue la valeur d'enchère afin de se rapprocher de la position ciblée. Après quelques itérations, ils réussissent à converger vers la valeur d'enchère qui leur permet d'atteindre la position souhaitée. À partir de ce moment, ils ne font plus varier leur valeur d'enchère, tant que la position moyenne demeure stable. Il est difficile de trouver des données historiques permettant d'effectuer des prédictions en fonction de la position lorsque les annonceurs utilisent une telle stratégie pour positionner leurs annonces, car l'étendue des données de position est souvent beaucoup trop restreinte.

Par ailleurs, même les annonceurs novices qui utilisent des stratégies de gestion moins complexes ont souvent tendance à retrouver leurs mots-clés dans des intervalles de position restreints. Certains ne visent même pas des positions particulières et se contentent de fixer leurs valeurs d'enchères à des seuils qu'ils considèrent comme acceptables. À cause de la grande taille de la majorité des campagnes, plusieurs d'entre eux ne modifient que très rarement leurs valeurs d'enchère pour un mot-clé donné. En effet, à moins d'utiliser des logiciels de gestion d'enchères, il est impossible de faire varier les valeurs d'enchère de plusieurs dizaines de milliers de mots-clés à chaque jour. Puisque la valeur d'enchère détermine ultimement la position des annonces, cette méthode de gestion génère des intervalles de positions très limités, ce qui a pour effet de rendre l'obtention de fonctions de prédiction plus difficile.

Conclusions

En considérant les problèmes soulevés lors de l'analyse préliminaire des données, nous arrivons à certaines conclusions concernant la nature des données de clics, CPC moyens et positions moyennes. Plus précisément, nous établissons la relation entre certaines caractéristiques des données et leur effet sur le potentiel de prédiction d'un mot-clé :

- Il est impossible d'obtenir des fonctions de prédiction qui modélisent bien l'ensemble des données lorsque la dispersion dans les données est trop grande.
- La qualité des fonctions de prédiction du nombre de clics et du CPC moyen en fonction de la position moyenne est généralement meilleure lorsque le nombre d'observations utilisées est relativement élevé.

- La qualité des fonctions de prédiction du nombre de clics et du CPC moyen en fonction de la position moyenne est généralement meilleure lorsque l'horizon temporel considéré pour la collecte de données n'est pas trop grand (i.e. les données utilisées ne sont pas trop anciennes).
- La qualité des fonctions de prédiction du nombre de clics en fonction de la position moyenne est généralement meilleure lorsque le nombre moyen de clics par observation est relativement élevé.
- La qualité des fonctions de prédiction du nombre de clics et du CPC moyen en fonction de la position moyenne est généralement meilleure lorsque la plage de positions dans laquelle se retrouvent les observations est plus large.

## 4.2 Algorithme de classification

En considérant les conclusions présentées à la section précédente, il n'est pas réaliste de penser qu'il soit possible de prédire le comportement de tous les mots-clés d'une campagne uniquement à l'aide de méthodes de régression. Dans certains cas, les caractéristiques des données historiques font en sorte que les régressions fournissent des prédictions de très mauvaise qualité. Il faudra donc procéder à une classification afin d'identifier les mots-clés qui ont un bon potentiel de prédiction avec les régressions. Par la suite, nous devrons également développer des solutions alternatives pour prédire le comportement de certains mots-clés, afin de maximiser l'étendue d'application du modèle d'optimisation globale.

D'abord, nous croyons qu'il est possible de définir des critères qui détermineront quels mots-clés d'une campagne peuvent être prédits adéquatement avec des méthodes de régression. Le Tableau 4.1 énumère les conditions qui, selon nous, devraient être respectées par un mot-clé pour qu'une régression puisse fournir des prédictions fiables à partir de ses données historiques. De plus, il fournit une mesure quantifiable permettant de mesurer objectivement le critère mentionné. Ces mesures pourraient être utilisées pour classer automatiquement les mots-clés en fonction de seuils prédéterminés ($\alpha$, $\beta$, $\gamma$, $\delta$ et $\varepsilon$). Les valeurs de seuils de classement réellement utilisées lors de nos analyses sont confidentielles et ne sont donc pas mentionnées. Elles ont été choisies suite à de

nombreuses analyses de données, de façon à obtenir un volume de mots-clés suffisant tout en assurant une bonne qualité au niveau des régressions.

Tableau 4.1 : Conditions nécessaires pour prédire le comportement d'un mot-clé à l'aide de régressions

| Condition | Mesure quantifiable |
|---|---|
| Nombre d'observations récentes suffisamment élevé | Nombre de jours de données disponibles dans les $\alpha$ derniers jours $\geq \beta$ |
| Nombre total de clics suffisamment élevé | Nombre total de clics $\geq \gamma$ |
| Faible dispersion dans les données | Coefficient de détermination $R^2$ (clics vs pos. moy.) $\geq \delta$ |
| | Coefficient de détermination $R^2$ (CPC moy. vs pos. moy.) $\geq \varepsilon$ |

Dans le cas des mots-clés qui ne respectent pas un ou plusieurs de ces critères, nous envisageons deux solutions qui permettraient d'obtenir des fonctions de prédiction fiables :

1) Modifier les valeurs d'enchère des mots-clés afin d'obtenir plus d'information concernant leur répartition de clics et de CPC moyens selon la position moyenne, dans le but d'appliquer des régressions à ces mots-clés par la suite (mécanisme de diversification des positions).

2) Développer des méthodes de prédiction alternatives afin d'estimer les fonctions de clics et de CPC moyen sans effectuer des régressions (fonctions de prédiction génériques).

La première solution est expliquée plus en détail à l'Annexe 2 et la seconde est présentée à la section 4.3. La Figure 4.5 illustre le rôle de chacune de ces solutions alternatives au sein de la structure de classification et leurs interactions avec les autres éléments de l'algorithme. On y retrouve tous les traitements et vérifications nécessaires pour arriver à un ensemble de mots-clés qui peut être optimisé par le modèle d'optimisation globale, incluant les conditions nécessaires pour l'obtention de régressions de qualité qui ont été mentionnées au Tableau 4.1.

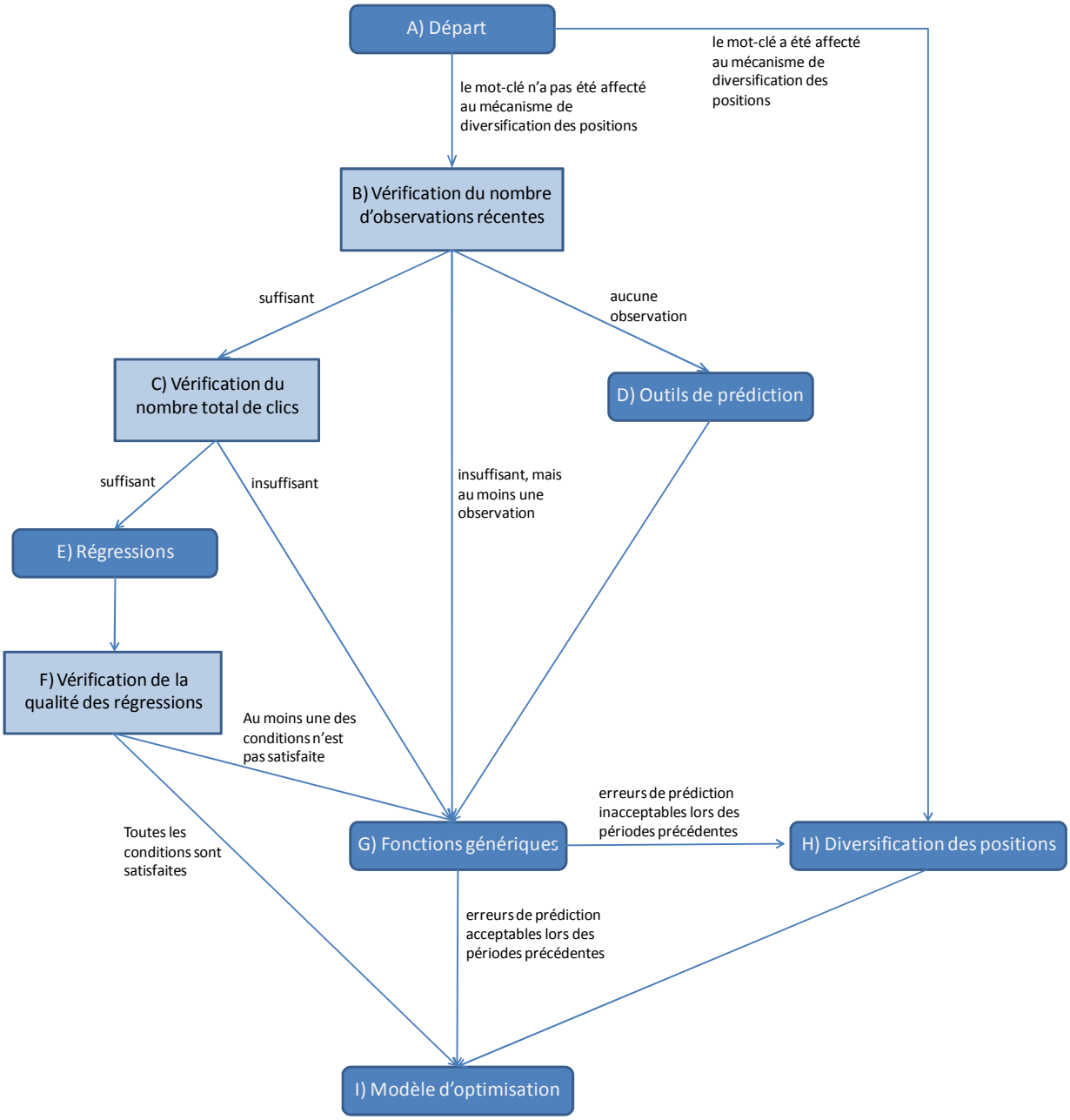



Figure 4.5 : Algorithme de classification

Cet algorithme de classification s'applique à chacun des mots-clés qui fait partie de l'ensemble à optimiser. En fonction de leurs caractéristiques et de la structure de leurs données historiques, les mots-clés seront affectés à un traitement ou une méthode de prédiction donnée. La stratégie envisagée consisterait à effectuer la classification des mots-clés avant chaque résolution du modèle d'optimisation, afin de s'assurer que les fonctions de prédiction fournissent des estimations fiables et que les solutions de l'optimisation soient utilisables. À mesure que des données historiques seront accumulées, il est possible que certains mots-clés soient classés différemment d'une période à l'autre, puisque leurs données historiques seront modifiées à mesure que de nouvelles observations seront recueillies. L'objectif ultime est d'arriver à une classification des mots-clés qui minimisera les erreurs de prédiction.

Explications détaillées

*A) Départ*

A est le point de départ pour chacun des mots-clés à l'étude. Tous les mots-clés qui font partie de l'ensemble à optimiser partiront de ce point de départ afin d'être classés.

*B) Vérification du nombre d'observations récentes*

Il faut vérifier le nombre d'observations récentes disponibles pour le mot-clé en question, afin de déterminer s'il sera possible de prédire son comportement avec des régressions. L'importance d'une quantité de données suffisante a été expliquée à la section 4.1.

*C) Vérification du nombre total de clics*

Il faut vérifier le nombre total de clics obtenus par le mot-clé en question lors de la période historique considérée, afin de déterminer s'il sera possible de prédire son comportement avec des régressions. L'importance d'un nombre de clics relativement élevé a été expliquée à la section 4.1.

*D) Outils de prédiction*

Les outils de prédiction sont fournis par les moteurs de recherche. Ils permettent d'estimer le rendement d'un mot-clé en fonction de la valeur d'enchère qui lui sera attribuée. Par exemple, l'estimateur de Google Adwords permet d'évaluer des valeurs approximatives du nombre de clics, CPC moyen et position moyenne qui seront obtenus pour un mot-clé donné, en fixant une valeur d'enchère précise et un budget quotidien. De plus, il donne une indication de la popularité du mot-clé en fournissant le nombre de recherches mensuelles qui lui sont associées. Même si ces estimations ne sont pas très précises, nous croyons que les outils de prédiction peuvent être utiles pour obtenir une estimation initiale du nombre de clics et des CPC moyens, dans le cas d'absence de données historiques.

*E) Régressions*

Une fois rendus à cette étape, des méthodes de régression sont appliquées aux mots-clés dans le but d'arriver à des fonctions de prédiction pour les clics et les CPC moyens. Dans le cadre de ce projet, ce sont des régressions linéaires et des régressions exponentielles linéarisées qui seront utilisées. Le choix de ces deux types de régression est basé sur les méthodes publiées dans la littérature ainsi que sur les analyses que nous avons effectuées à partir de nos banques de données. D'abord, plusieurs auteurs du domaine prétendent utiliser des fonctions exponentielles dans leurs modèles. Entre autres, Kitts & Leblanc (2004) ainsi que Kitts et al. (2005) obtiennent leurs fonctions de clics avec des régressions exponentielles. De plus, Kitts & Leblanc (2004) et Ganchev et al. (2007) expliquent que les CPC moyens décroissent de façon exponentielle selon leurs analyses. Par ailleurs, même si les références dans la littérature semblent pencher vers l'utilisation des fonctions exponentielles, nos analyses ont démontré que les régressions linéaires permettent aussi de modéliser adéquatement une bonne proportion des mots-clés. Nous croyons donc qu'il serait pertinent d'appliquer les deux types de régression et de conserver celui qui s'ajuste le mieux aux données du mot-clé considéré.

Nous prévoyons possiblement ajouter d'autres types de régressions à cette étape dans le futur. Nous sommes conscients qu'il existe plusieurs méthodes de régression qui pourraient potentiellement fournir de meilleures prédictions que les régressions linéaire simple et exponentielle linéarisée simple. Par exemple, il serait possible d'utiliser des régressions à

variables multiples (considérer l'effet d'autres facteurs sur la prédiction du nombre de clics et des CPC moyens), des régressions pondérées, des régressions exponentielles non linéarisées (méthode des moindres carrés non linéaires qui procède de façon itérative pour trouver un ajustement), des régressions combinées à des séries temporelles, etc. Cependant, puisque ces méthodes n'ont pas été testées dans le cadre de ce mémoire, elles ne sont pas abordées.

*F) Vérification de la qualité des régressions*

Pour que nous considérions qu'un mot-clé puisse être prédit à l'aide de méthodes de régression, il est nécessaire que ses deux fonctions de prédiction (clics et CPC moyen en fonction de la position moyenne) satisfassent certaines conditions de qualité. Entre autres, il faut évaluer la qualité des fonctions de régression selon certains critères statistiques.

Nous proposons d'abord d'effectuer l'analyse des résidus. En effet, il est nécessaire de vérifier que les erreurs suivent une distribution normale et sont indépendamment distribuées avant de pouvoir appliquer le test de signification de la régression. Il serait possible d'automatiser l'analyse des résidus d'une régression en calculant la valeur $W$ du test de Shapiro-Wilk et en la comparant à une valeur critique qui peut être obtenue dans des tables de Shapiro-Wilk. Si la valeur calculée est inférieure au seuil déterminé, nous pouvons supposer la normalité des résidus.

Dans les cas où l'analyse des résidus révèle que les erreurs sont normales et indépendamment distribuées, il est possible de poursuivre avec le test d'hypothèses de signification de la régression. Ce test implique simplement de calculer une valeur $t_0$ et la comparer à une valeur critique $t_{\alpha/2,n-2}$ donnée dans la table de la loi de Student, qui dépend du niveau de confiance exigé $\alpha$ et du nombre d'observations $n$. Il serait possible d'automatiser le calcul de ce test d'hypothèses dans le but de l'appliquer à plusieurs milliers de mots-clés.

$$t_0 = \frac{\widehat{\beta_1}}{\sqrt{MS_E/S_{xx}}}$$

où $\widehat{\beta_1}$ est la pente estimée par la régression, $MS_E$ est la moyenne du carré des erreurs sur les observations de clics ou de CPC moyen par rapport aux régressions et $S_{xx}$ est la somme du carré des écarts des observations en $x$ (position moyenne) par rapport à la valeur moyenne $\bar{x}$. [3]

Si $|t_0| > t_{\alpha/2,n-2}$, il est possible de conclure que la régression est significative. Plus précisément, cela confirme que la variable indépendante (position moyenne) exerce une influence significative sur la variable dépendante (nombre de clics ou CPC moyen). Dans ce cas, nous évaluons de façon plus précise la qualité de la fonction de régression en utilisant son coefficient de détermination $R^2$ comme critère de qualité. Les coefficients de détermination associés aux régressions doivent excéder un seuil prédéterminé (par exemple, $R^2 > 0{,}30$) pour que la régression soit retenue.

Remarque : Le coefficient de détermination ($R^2$) est une mesure fréquemment utilisée pour juger la qualité d'ajustement d'un modèle de régression. Il s'agit d'une valeur entre 0 et 1 qui représente « la quantité de variabilité dans les données expliquée ou justifiée par le modèle de régression » (Hines et al., 2005, p.382). Par conséquent, il fournit souvent une mesure de la précision avec laquelle les observations futures risquent d'être prédites. Cependant, il faut être vigilant avec l'utilisation de cette mesure, car une bonne valeur de $R^2$ n'implique pas nécessairement que le modèle est juste.

Par la suite, nous évaluons les fonctions obtenues d'un point de vue logique. Il faut qu'elles respectent certaines conditions nécessaires pour assurer une cohérence au niveau de la modélisation :

- La fonction de régression des clics décroît lorsque la position augmente.
- La fonction de régression des CPC moyens décroît lorsque la position augmente.
- Les valeurs prédites par la fonction de régression des clics sont positives pour au moins les $n$ premières positions (où $n$ est un paramètre à définir)

[3] Pour plus de détails, voir Hines et al. (2005), section 13.2

- Les valeurs prédites par la fonction de régression des CPC moyens sont positives pour au moins les $m$ premières positions (où $m$ est un paramètre à définir)

Ces conditions vérifient essentiellement si les régressions fournissent des prédictions conformes aux hypothèses sur lesquelles notre modèle d'optimisation est basé. Par exemple, une fonction qui estimerait des valeurs de CPC moyen croissantes en fonction de la position serait inacceptable, car nous savons que les CPC moyens doivent généralement diminuer lorsque la position augmente. La non-satisfaction d'une de ces conditions aurait pour effet d'affecter sérieusement la qualité des solutions fournies par le modèle d'optimisation.

Les régressions qui satisfont le seuil de $R^2$ minimal et les conditions nécessaires seront acceptées et la fonction de prédiction obtenue sera utilisée dans le modèle d'optimisation (*I*). Dans le cas des régressions qui ne satisfont pas le seuil de $R^2$ minimal ou les conditions nécessaires, nous rejetterons la régression et utiliserons des fonctions génériques (*G*) pour obtenir une fonction de prédiction. Notons qu'il est possible qu'une seule des deux fonctions de prédiction d'un mot-clé (clics ou CPC moyen) soit rejetée. Dans ce cas, les fonctions génériques seront utilisées uniquement pour prédire la fonction rejetée. L'autre fonction sera prédite à l'aide d'une régression.

Pour chaque mot-clé, nous proposons d'appliquer tout le processus de vérification mentionné ci-haut (analyse des résidus, test d'hypothèse de signification de la régression, calcul du $R^2$ et vérification des conditions logiques) aux deux types de régressions que nous considérons : la régression linéaire et la régression exponentielle linéarisée. Dans le cas où les deux régressions satisfont toutes les exigences, il serait préférable de prendre celle qui offre la meilleure valeur de $R^2$. Si une seule des deux régressions satisfait toutes les conditions, le choix est alors trivial. Finalement, si aucune des deux régressions ne satisfait toutes les conditions, le comportement du mot-clé ne peut être prédit avec une régression et il faut utiliser une méthode de prédiction alternative (les fonctions génériques) pour obtenir des estimations. Notons que pour un mot-clé donné, il est possible que le type de fonction utilisé pour la prédiction des clics ne soit pas le même que celui utilisé pour la prédiction des CPC moyens.

Pour terminer, il faut apporter certaines modifications aux méthodes de prédiction pour les adapter aux réalités du problème. Entre autres, il faut éliminer les prédictions négatives, car il est impossible d'obtenir un nombre de clics ou un CPC moyen négatif. Pour y arriver, nous

définissons des variables de position uniquement pour les positions qui offrent des prédictions non négatives. Ainsi, la position maximale ($Pos_{max}$) considérée pour chaque mot-clé sera le minimum entre la position maximale de 10 mentionnée à la section 3.2 et la position minimale pour laquelle les clics prédits ($Pos_{max}^{clics}$) et CPC moyens prédits ($Pos_{max}^{cpc}$) sont positifs ou nuls :

$$Pos_{max} = min\{10, Pos_{max}^{clics}, Pos_{max}^{cpc}\}$$

Le modèle ne pourra donc pas prédire une valeur de clics ou de CPC moyen négative à partir des régressions obtenues.

Remarque : En imposant des conditions relatives à l'analyse des résidus et au test de signification de la régression, nous nous assurons que notre démarche soit mathématiquement acceptable. Cependant, il est possible que ces conditions réduisent considérablement le nombre de mots-clés qui peuvent être prédits avec des régressions. Il faudra analyser plus en profondeur pour déterminer si ces conditions devraient être conservées, modifiées ou éliminées.

*G) Fonctions génériques*

Les fonctions génériques sont expliquées en détail au Chapitre 6. Il s'agit essentiellement de méthodes de prédiction qui approximent les taux de décroissance des variables de clics et de CPC moyen d'un mot-clé donné en se basant sur les données d'un ensemble de mots-clés semblables. Puisque les fonctions de prédiction génériques ne dépendent pas de la qualité des données historiques actuelles du mot-clé, on peut les appliquer lorsque les régressions ne sont pas acceptées ou lorsque les conditions nécessaires ne sont pas vérifiées.

Il suffit d'une observation pour appliquer les taux de décroissance génériques à chaque mot-clé afin d'obtenir une fonction générique. C'est pourquoi, dans les cas où il n'existe aucune observation récente dans la base de données, les outils de prédiction (*D*) sont utilisés afin de fournir un point de départ aux fonctions génériques.

Évidemment, une méthode de prédiction générique risque de ne pas toujours prédire avec exactitude. Dans les cas où les erreurs de prédiction moyennes sont trop élevées (moyenne calculée sur plusieurs jours passés), nous concluons que les fonctions génériques ne sont pas

applicables et le mot-clé est envoyé au mécanisme de diversification des positions (*H*). Lorsque les erreurs moyennes n'ont pas encore été comptabilisées ou sont considérées comme acceptables, les fonctions génériques sont utilisées pour obtenir les fonctions de prédiction du modèle d'optimisation (*I*).

*H) Mécanisme de diversification des positions*

Le mécanisme de diversification des positions est expliqué plus en détail à l'Annexe 2. Il s'agit d'un algorithme qui fait varier les valeurs d'enchère d'un mot-clé afin d'acquérir un maximum d'information concernant ses fonctions de clics et de CPC moyen. Plus précisément, il vise à occuper des positions non visitées ou rarement visitées par un mot-clé, dans le but de préciser les fonctions sur une plus large étendue de positions. Tel qu'expliqué à la section 4.1, le phénomène de faible étendue de positions est très fréquemment rencontré et la variation des enchères permet de maximiser le nombre de positions différentes qui sont occupées. Dans plusieurs cas, ceci permet d'améliorer considérablement la qualité des fonctions de régression obtenues.

Lors du passage d'un mot-clé dans l'étape *H* de l'algorithme de classification, il faudra déterminer la durée pour laquelle il sera géré par le mécanisme de diversification des positions. De cette façon, au début de la classification (*A*), chaque mot-clé affecté temporairement à la diversification des positions sera directement dirigé vers l'étape *H*. À chaque jour de diversification, ce sera le mécanisme qui déterminera la valeur d'enchère du mot-clé en question. Afin de considérer cette valeur d'enchère et la soustraire au budget global quotidien, nous proposons de fixer à 1 la valeur de la variable de position $y_{mp}$ correspondante. Par exemple, si un mot-clé $m$ est affecté à la position 8, il faudra fixer $y_{m8} = 1$ pour ce mot-clé dans le modèle d'optimisation (*I*). Suite à la diversification des positions, il faudra appliquer les méthodes de régression aux données obtenues pour déterminer si les modifications ont permis d'augmenter la qualité des prédictions.

*I) Modèle d'optimisation*

Le modèle d'optimisation a été décrit au Chapitre 4. L'objectif de l'algorithme de classification est de fournir des prédictions fiables à ce modèle d'optimisation, afin que les solutions qu'il obtient permettent de réellement améliorer le rendement des campagnes publicitaires.

Répartition actuelle des mots-clés

En effectuant des essais de classification sur certaines banques de données à notre disposition, nous sommes arrivés à quelques observations intéressantes concernant la répartition des mots-clés dans chacune des sections de cet algorithme de classification. D'abord, nous avons pu constater que le nombre de mots-clés qui satisfont les critères de nombre d'observations récentes (*B*) et nombre total de clics (*C*) sont assez limités. Cependant, ceux qui satisfont ces deux critères représentent habituellement une proportion considérable du volume et des coûts de la campagne publicitaire concernée. Il peut donc être profitable de réussir à prédire le comportement de ces mots-clés malgré leur faible quantité.

À cause des méthodes de gestion de campagnes généralement utilisées par les annonceurs, les valeurs d'enchère demeurent relativement fixes et il est souvent difficile d'obtenir des fonctions de régression fiables pour les clics et les CPC moyens. C'est pourquoi le nombre de mots-clés qui satisfont tous les critères de qualité à l'étape *F* sont très rares, lorsque nous effectuons des tests sur les données des campagnes publicitaires actuellement à notre disposition. Toutefois, ce problème ne devrait pas persister lors de l'utilisation périodique du modèle d'optimisation globale; celui-ci risque de faire varier les valeurs d'enchère, ce qui aura pour effet de fournir de meilleures régressions dans le futur. De plus, les fonctions génériques et le mécanisme de diversification des positions sont deux solutions alternatives qui permettraient de prédire le comportement des mots-clés avec lesquels il est initialement impossible d'obtenir des régressions de qualité, donc il sera toujours possible d'effectuer des prédictions plus ou moins fiables pour l'ensemble des mots-clés. Bref, les mots-clés pouvant être prédits à l'aide de régressions sont présentement très rares, mais nous sommes confiants que la variation des valeurs d'enchère engendrée par l'utilisation du modèle d'optimisation (*I*) en parallèle avec les fonctions génériques (*G*) ou le mécanisme de diversification des positions (*H*) aura pour effet d'améliorer le potentiel de prédiction des mots-clés avec le temps.

Remarque

Initialement, nous avions l'intention d'affecter au mécanisme de diversification tous les mots-clés pour lesquels les régressions ne fournissaient pas des prédictions acceptables ainsi que tous ceux qui ne possédaient pas suffisamment de données historiques. Cependant, nous avons rapidement constaté que cela engendrerait probablement une diminution du rendement de la campagne. En effet, en faisant constamment varier les enchères d'un mot-clé afin de maximiser le nombre de positions différentes visitées, nous risquons de nous éloigner de certaines positions performantes pendant une période de temps de durée non négligeable. Cette expérimentation permet d'acquérir de l'information nouvelle, mais il faut attendre un certain temps et débourser des sommes considérables avant d'obtenir toutes les données nécessaires. Bref, la diversification des positions est une solution fonctionnelle, mais qui comporte certains inconvénients.

Nous avons donc décidé d'orienter nos efforts à développer une méthode de prédiction alternative plus rapide et moins coûteuse. C'est à ce moment que nous avons approfondi l'idée des fonctions génériques de prédiction. À notre avis, il s'agit d'une méthode beaucoup plus avantageuse pour prédire le comportement des mots-clés qui ne fournissent pas de bonnes prédictions avec les régressions. C'est pourquoi le mécanisme de diversification des positions est maintenant réservé uniquement aux mots-clés qui ne fournissent pas des prédictions acceptables avec les fonctions génériques. La majorité de nos efforts de recherche ont été consacrés au développement et à l'amélioration de nos fonctions de prédiction génériques. Puisque les prédictions obtenues par ces fonctions sont relativement bonnes pour l'ensemble des campagnes publicitaires testées jusqu'à présent, nous considérons même possiblement d'éliminer l'utilisation du mécanisme de diversification. Bref, ces raisons justifient pourquoi un chapitre entier est nécessaire pour la présentation des fonctions de prédiction génériques, tandis qu'une brève description en annexe est suffisante pour l'explication du mécanisme de diversification.

# CHAPITRE 5 FONCTIONS DE PRÉDICTION GÉNÉRIQUES

Tel qu'expliqué dans les chapitres précédents, l'obtention de régression fiables n'est pas toujours possible lors de la prédiction des clics et des CPC moyens en fonction de la position moyenne. Des facteurs tels qu'une insuffisance ou absence de données historiques, un nombre moyen de clics par observation trop faible et une plage de données de position trop étroite peuvent faire en sorte que l'effet de décroissance ne puisse pas être observé.

Même si cet effet n'est pas toujours observé à partir des données disponibles, nous savons tout de même que dans la majorité des cas, la position moyenne a une influence sur le nombre de clics et le CPC moyen. Comme nous l'avons mentionné à la section 1.4, les études effectuées par les entreprises impliquées dans le domaine, les références publiées dans la littérature ainsi que nos propres études démontrent toutes que les taux de clic décroissent généralement lorsque la position baisse. De plus, l'algorithme de classement des annonces expliqué à la section 1.5 est défini de façon à ce que les CPC associés à chacune des positions soient décroissants, ce qui fait en sorte que les CPC moyens agrégés quotidiennement sont presque toujours décroissants en fonction de la position. Bref, il existe très fréquemment une relation décroissante entre le nombre de clics obtenus par un mot-clé et sa position moyenne, ainsi qu'entre le CPC moyen d'un mot-clé et sa position moyenne. Nous cherchons donc une façon d'estimer le comportement d'un mot-clé lorsque ses données historiques ne nous permettent pas de générer des fonctions de prédiction de façon exacte.

## 5.1 Principe

Sachant que le nombre de clics et le CPC moyen sont des variables décroissantes en fonction de la position moyenne, nous sommes portés à nous demander si la décroissance relative de ces fonctions est sensiblement la même d'un mot-clé à l'autre. Plus précisément, nous cherchons à déterminer si les fonctions de prédiction sont les mêmes d'un mot-clé à l'autre lorsque leurs valeurs de variables dépendantes sont ajustées à une échelle commune. Dans l'affirmative, il serait possible d'utiliser des fonctions à taux de décroissance uniformes pour estimer le comportement de tous les mots-clés. Nous appellerons celles-ci des « fonctions génériques », car elles permettraient de prédire un ensemble de mots-clés à partir d'un même taux de décroissance.

Peu importe la requête qui est effectuée, les emplacements réservés aux annonces textuelles sont toujours situés aux mêmes endroits sur les pages de recherche. Évidemment, le nombre d'annonces peut varier d'une requête à l'autre et les annonces peuvent se déplacer d'une position à l'autre lorsque les valeurs d'enchère de leurs mots-clés sont modifiées, mais l'emplacement de chacune des positions potentielles demeure standard. Sachant cela, il est raisonnable de croire que les clics générés par les utilisateurs se répartissent proportionnellement de la même façon entre chacune des positions d'annonces potentielles, peu importe le mot-clé. Dans un tel cas, les clics obtenus par les mots-clés suivraient tous la même décroissance relative et il serait possible de les prédire à l'aide de fonctions génériques. Par ailleurs, si les clics obtenus à chacune des positions suivent une répartition relativement constante, il est également possible que les valeurs de CPC moyen suivent un taux de décroissance presque constant d'un mot-clé à l'autre. Si cela se confirmait, il serait possible d'utiliser des fonctions génériques pour estimer les CPC moyens. Bref, nous cherchons à confirmer les hypothèses selon lesquelles les clics et les CPC moyens ont les mêmes décroissances relatives pour tous les mots-clés, afin de déterminer si l'utilisation de fonctions de prédiction génériques est possible.

Tel que mentionné au chapitre précédent, nous avons constaté que les régressions linéaires et exponentielles fonctionnent bien pour prédire les clics et les CPC moyens selon la position. Nous avons donc orienté nos recherches de fonctions génériques autour de ces deux types de fonctions. La section 5.2 présente d'abord les données utilisées pour effectuer les analyses qui sont présentées tout au long de ce chapitre. Par la suite, la section 5.3 explique nos démarches pour tester la prédiction à l'aide de fonctions linéaires, puis la section 5.4 présente les analyses effectuées avec les fonctions exponentielles. À la section 5.5, nous comparons les résultats obtenus avec chacune de ces méthodes afin de choisir la meilleure approche, autant pour la prédiction des clics que pour la prédiction des CPC moyens. Finalement, à la section 5.6, nous appliquons quelques raffinements à la méthode choisie afin d'améliorer son rendement et nous discutons du potentiel d'intégration de cette méthode de prédiction avec le modèle d'optimisation.

## 5.2 Données utilisées pour les analyses

Banques de données à tester

Afin de valider nos hypothèses avec un échantillon représentatif de données, il était important de sélectionner plusieurs banques de données suffisamment diversifiées pour effectuer nos analyses. En effet, nous voulions sélectionner des ensembles de mots-clés provenant de plusieurs types d'annonceurs différents, afin de vérifier si les conclusions de l'étude s'appliquent dans tous les domaines, pour tous les types de produits ou services offerts. De plus, nous devions nous assurer que ces banques de données totalisaient des volumes significatifs, autant au niveau du nombre de mots-clés et de journées historiques disponibles que pour les volumes de clics, d'impressions et de coûts.

Nous avons donc sélectionné un ensemble de 20 banques de données provenant de plusieurs types de marchés différents. Pour des raisons de confidentialité, les annonceurs associés à chacune de ces banques de données ne sont pas mentionnés. La majorité d'entre elles possèdent des volumes très élevés, mais nous en avons également inclus quelques-unes avec des volumes plus faibles, afin de vérifier si leurs comportements sont différents. Les caractéristiques de ces banques de données sont présentées au Tableau 5.1. Avec une telle diversité et des volumes totaux aussi importants, nous considérons que les conclusions atteintes à partir de cet ensemble seront applicables à la majorité des campagnes publicitaires qui sont gérées par la plateforme logicielle d'Acquisio.

Tableau 5.1 : Description globale des banques de données utilisées pour les analyses

| Compte | nombre de jours | nombre total de mots-clés | nombre total d'impressions | nombre total de clics | somme des coûts ($) |
|---|---|---|---|---|---|
| 1 | 1483 | 271500 | 844454748 | 20197449 | 19017389 |
| 2 | 1798 | 319268 | 1123858823 | 17455366 | 15440329 |
| 3 | 1450 | 91836 | 839032278 | 8061347 | 5884004 |
| 4 | 1160 | 16835 | 189657547 | 2386089 | 1302552 |
| 5 | 763 | 45685 | 518932623 | 10071809 | 5182216 |
| 6 | 690 | 4875 | 45005017 | 396310 | 549224 |
| 7 | 1084 | 15799 | 268256181 | 2802682 | 2559622 |
| 8 | 390 | 5259 | 24220323 | 509820 | 408544 |
| 9 | 442 | 14262 | 380276817 | 2456660 | 2394476 |
| 10 | 824 | 2964 | 25037337 | 444914 | 226411 |
| 11 | 398 | 36027 | 57588673 | 672728 | 416652 |
| 12 | 260 | 570 | 2093065 | 47030 | 23706 |
| 13 | 270 | 3667 | 8835208 | 320063 | 102876 |
| 14 | 224 | 2789 | 81064428 | 122860 | 113237 |
| 15 | 74 | 5589 | 57851978 | 38490 | 54541 |
| 16 | 118 | 20494 | 36870034 | 420608 | 748448 |
| 17 | 1613 | 80898 | 2234789491 | 97882954 | 15140767 |
| 18 | 571 | 138400 | 634306754 | 5129734 | 10845912 |
| 19 | 1842 | 310815 | 1705546971 | 72568128 | 34847682 |
| 20 | 1377 | 23855 | 373573967 | 3475079 | 2498015 |
| | somme | 1411387 | 9451252263 | 245460120 | 117756603 |

Répartition des clics et des coûts dans une campagne publicitaire

Nous jugeons qu'il est important de comprendre la structure des campagnes publicitaires étudiées avant de procéder à des analyses de données. Afin d'observer la répartition des volumes de clics et de coûts pour chacune des banques de données, nous avons mis en évidence la somme des clics moyens par jour et la somme des coûts moyens par jour en fonction des mots-clés, en ordonnant ces mots-clés en ordre décroissant de volume. Nous avons choisi d'utiliser des valeurs moyennes par jour plutôt que des valeurs absolues afin d'éliminer le biais associé au fait que les mots-clés n'ont pas tous le même nombre de jours de données historiques. Pour chacune des banques de données, nous avons créé 20 groupes de mots-clés : le premier groupe représente les mots-clés qui se classent du 95$^{e}$ centile au 100$^{e}$ centile, le deuxième groupe représente ceux qui se classent du 90$^{e}$ centile au 95$^{e}$ centile et ainsi de suite (dépendant du tableau, les rangs centiles sont déterminés par le nombre de clics moyen par jour ou les coûts moyens par jour associés aux mots-clés). Les résultats sont présentés sous forme de tableaux à l'Annexe 3. Le but de cette classification est de démontrer à quel point le volume des campagnes publicitaires est concentré dans un nombre limité de mots-clés.

Le Tableau 5.2 montre cette répartition des clics et des coûts moyens par jour obtenus lorsque nous agrégeons les résultats des 20 banques de données présentées à l'Annexe 3. Ces résultats nous permettent de constater qu'une très faible proportion des mots-clés est responsable pour la forte majorité des clics et des coûts. Plus précisément, les résultats agrégés démontrent qu'il suffit de 5% des mots-clés pour totaliser 88,34% du total des clics moyens par jour et 87,48% des coûts moyens par jour. De plus, au moins 35% des mots-clés ne génèrent absolument aucun clic et, par conséquent, aucun coût.

Tableau 5.2 : Répartition des clics et des coûts en fonction des mots-clés pour toutes les banques de données agrégées

| | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 1 119 916,61 | 88,34% | 835512,46 | 87,48% |
| 90-95 | 64 126,42 | 5,06% | 57427,72 | 6,01% |
| 85-90 | 29 971,75 | 2,36% | 25818,71 | 2,70% |
| 80-85 | 18 303,12 | 1,44% | 14589,29 | 1,53% |
| 75-80 | 12 362,75 | 0,98% | 8931,40 | 0,94% |
| 70-75 | 8 600,21 | 0,68% | 5470,63 | 0,57% |
| 65-70 | 5 953,72 | 0,47% | 3320,07 | 0,35% |
| 60-65 | 4 042,11 | 0,32% | 2089,28 | 0,22% |
| 55-60 | 2 388,00 | 0,19% | 1116,96 | 0,12% |
| 50-55 | 1 227,18 | 0,10% | 537,27 | 0,06% |
| 45-50 | 599,72 | 0,05% | 229,02 | 0,02% |
| 40-45 | 198,22 | 0,02% | 70,99 | 0,01% |
| 35-40 | 18,48 | 0,00% | 7,29 | 0,00% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

Cette structure est très courante dans le domaine du marketing sur les moteurs de recherche; il est typique d'avoir un nombre très élevé de mots-clés à faible volume et un nombre très limité de mots-clés à haut volume. Devant ces faits, nous concluons qu'il suffit d'optimiser le rendement d'un petit sous-ensemble de mots-clés pour améliorer considérablement la performance globale d'une campagne publicitaire. Les stratégies d'optimisation et de prédiction utilisées devraient donc être élaborées en tenant compte de cette répartition des clics et des coûts au sein des campagnes publicitaires.

## 5.3 Fonctions génériques linéaires

Ajustement d'échelle

Pour arriver à comparer les taux de décroissance relatifs des mots-clés, il faut d'abord convertir les données de variables dépendantes de ces mots-clés à une échelle commune. Afin d'obtenir cette échelle commune, il suffit de pondérer les nombres de clics et CPC moyens par une valeur qui caractérise l'échelle de la variable pour le mot-clé en question. Nous choisissons donc de réduire l'échelle des graphiques en divisant chacune des valeurs de clics par le nombre de clics prédit en position 1 et en divisant chacune des valeurs de CPC moyen par le CPC moyen prédit en position 1 (les valeurs prédites sont obtenues en appliquant des régressions linéaires aux données historiques). Par la suite, il est possible de comparer les pentes des régressions mises à l'échelle.

La Figure 5.1 montre l'exemple de deux mots-clés qui ont sensiblement la même décroissance linéaire relative pour leur fonction de clics. Dans cet exemple, le premier mot-clé (graphique A) a un volume de clics beaucoup moins élevé que le second (graphique B). Cependant, lorsque nous effectuons une réduction d'échelle sur ces données et leur appliquons une régression linéaire (graphiques C et D), nous constatons que les nouvelles pentes obtenues sont presque les mêmes, soit -0,117 pour le premier mot-clé et -0,120 pour le second. Notons que ces valeurs auraient également pu être obtenues en divisant les pentes des graphiques initiaux (A et B) par leurs valeurs prédites en position 1. Puisque les valeurs des pentes mises à l'échelle sont presque les mêmes, nous en déduisons que les deux mots-clés ont des répartitions de clics semblables.

Extraction d'ensembles de mots-clés à tester

Notre objectif consiste donc à déterminer s'il existe des valeurs constantes vers lesquelles tendent les pentes des graphiques de clics et CPC moyens mis à l'échelle, pour chacun des mots-clés à l'étude. Pour y arriver, il est nécessaire d'effectuer des analyses uniquement à partir de mots-clés qui génèrent des régressions linéaires acceptables pour leurs fonctions de clics et de CPC moyens. En effet, tout comme dans l'exemple précédent, nous devrons calculer les régressions d'un ensemble de mots-clés, puis comparer les valeurs de pentes mises à l'échelle commune afin de déterminer si elles ont des valeurs semblables.

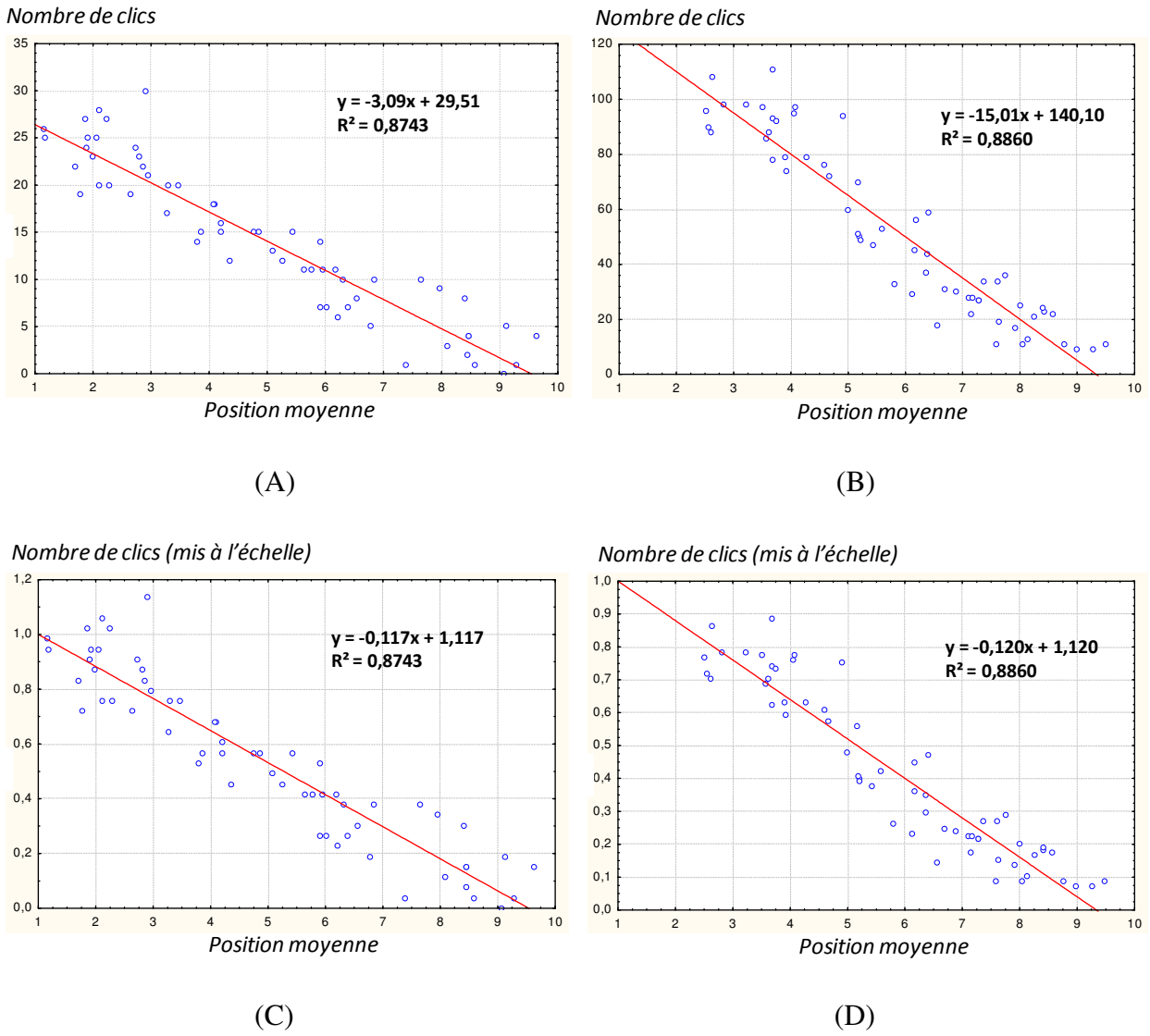



Figure 5.1 : Exemples de graphiques de clics pour deux mots-clés avec taux de décroissance relatifs semblables, avant (A et B) et après (C et D) mise à l'échelle

Dans cette optique, nous avons défini des critères nous permettant d'extraire, à partir des 20 banques de données disponibles pour l'analyse, un ensemble de mots-clés offrant des fonctions de prédiction de qualité. En imposant des conditions très précises, nous avons obtenu 206 mots-clés qui génèrent des fonctions de régressions linéaires que nous considérons comme acceptables pour la prédiction des clics et 184 mots-clés pour la prédiction des CPC moyens. Voici les critères qui ont été utilisés pour obtenir ces mots-clés :

- écart-type des valeurs de position $\geq 1,5$
- position maximale - position minimale $\geq 4$
- nombre de jours de données disponibles dans les 120 derniers jours $\geq 100$
- nombre moyen de clics par jour $\geq 20$
- valeur de position minimale $\leq 5$ (i.e. nous exigeons qu'au moins une observation se situe entre les positions 1 et 5 inclusivement)
- coefficient de détermination ($R^2$) de la régression linéaire $\geq 0,30$
- la fonction de régression des clics décroît lorsque la position augmente (pente négative)
- la fonction de régression des CPC moyens décroît lorsque la position augmente (pente négative)
- les valeurs prédites par la fonction de régression des clics sont positives pour au moins les 5 premières positions
- les valeurs prédites par la fonction de régression des CPC moyens sont positives pour au moins les 5 premières positions

Nous considérons que les mots-clés qui respectent toutes ces conditions risquent fort probablement de fournir des régressions fiables. Par conséquent, ils constituent de bons candidats pour vérifier les hypothèses selon lesquelles les mots-clés suivent tous sensiblement les mêmes taux de décroissance relatifs. Même si les fonctions de prédiction génériques sont destinées à être utilisées avec les mots-clés dont les données historiques ne permettent pas l'utilisation des méthodes de régression, il faut d'abord les valider avec des mots-clés fournissant des régressions de qualité.

*Remarque* : Pour assurer la qualité des régressions obtenues, il faudrait idéalement vérifier la normalité des résidus et effectuer un test d'hypothèse sur la signification de la régression (tel qu'expliqué à la section 4.2). Cependant, ces procédures de vérification n'ont pas été programmées dans le cadre de ce mémoire, donc nous nous contentons des conditions mentionnées ci-haut. Nous croyons qu'avec nos exigences sévères au niveau de la répartition des

positions occupées, nous pouvons éliminer la majorité des cas extrêmes qui fourniraient des régressions inacceptables.

Évidemment, le nombre de mots-clés obtenu pour chacun des deux ensembles à tester (206 et 184) est très faible par rapport au nombre total de mots-clés contenus dans les 20 banques de données. Cela peut être expliqué par le fait qu'il existe très peu de mots-clés à haut volume (au moins 20 clics par jour en moyenne) et encore moins de mots-clés qui visitent une grande plage de positions (écart-type des valeurs de position $\geq 1,5$ et position maximale – position minimale $\geq 4$). En imposant ces conditions ainsi que toutes les autres mentionnées, il est normal que le nombre de mots-clés obtenu soit très limité. Nous sommes conscients que les critères utilisés sont très sévères, mais ils sont nécessaires puisque nous désirons travailler avec des mots-clés qui fournissent des régressions acceptables. En diminuant nos exigences, les ensembles de mots-clés obtenus seraient plus grands, mais il ne serait pas toujours possible de se fier aux régressions obtenues pour estimer les taux de décroissance des mots-clés. Afin de valider si les conclusions obtenues à partir de ces sous-ensembles de mots-clés s'appliquent réellement à tous les autres mots-clés d'une campagne publicitaire, nous procédons à des analyses d'erreur très détaillées aux sections 5.5 et 5.6.

Analyse des pentes mises à l'échelle pour les graphiques de clics

Pour vérifier si les pentes des graphiques de clics mis à l'échelle convergent vers une valeur constante, il suffit de calculer chacune de leurs régressions et comparer les valeurs de pentes obtenues. Puisque l'ensemble à tester pour la prédiction des clics est constitué de 206 mots-clés, nous devons effectuer 206 régressions linéaires (tout comme ce qui a été fait pour l'exemple de la Figure 5.1 C et D). La Figure 5.2 illustre la répartition des pentes obtenues pour chacune de ces régressions. Afin de fournir une meilleure vue globale sur l'ensemble du nuage de points, deux valeurs extrêmes ont été enlevées du graphique. Nous justifions la suppression de ces valeurs extrêmes par le fait qu'elles se retrouvaient à l'extérieur des bornes *[médiane-3*étendue interquartile, médiane+3*étendue interquartile]*, l'étendue interquartile étant calculée par la différence entre la valeur du 3$^{e}$ quartile et celle du 1$^{er}$ quartile. Suite à cette suppression, il nous reste donc 204 observations sur le graphique plutôt que 206.

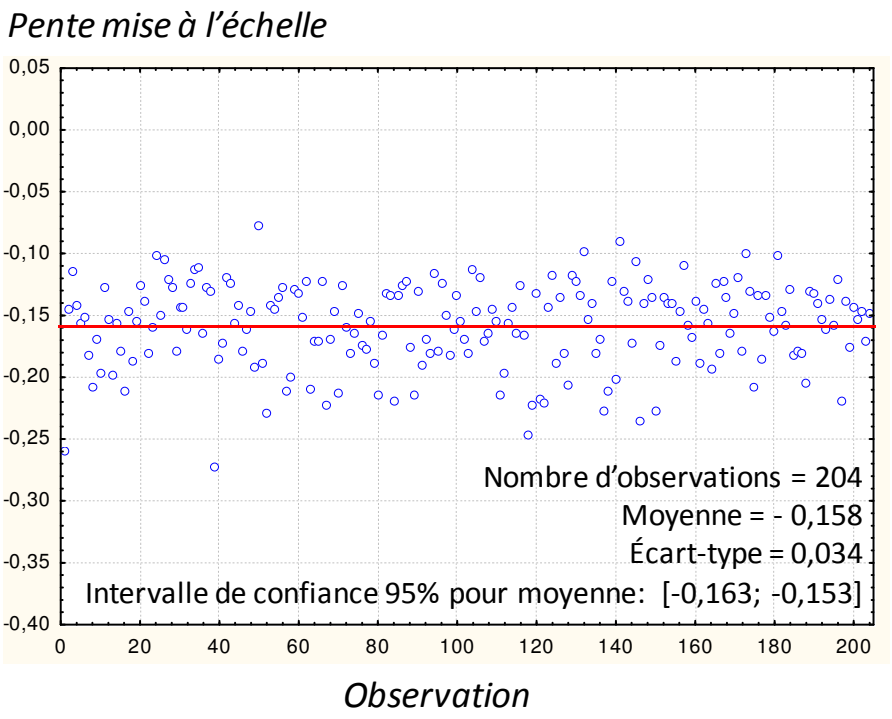



Figure 5.2 : Répartition des pentes mises à l'échelle pour les graphiques de clics en fonction de la position moyenne

En observant cette figure, nous constatons que les valeurs de pentes mises à l'échelle se regroupent autour d'une moyenne d'environ -0,158. Évidemment, les pentes ne sont pas toutes exactement les mêmes et il existe une certaine variation entre chacune des observations. Cependant, les valeurs semblent tendre vers la moyenne, car l'écart-type n'est que de 0,034 et visuellement, les points sont relativement peu distancés. En considérant les nombreux facteurs pouvant provoquer des fluctuations aléatoires dans les données, nous croyons qu'une telle répartition permet d'affirmer que les taux de décroissance des graphiques mis à l'échelle sont sensiblement les mêmes d'un mot-clé à l'autre.

Analyse des pentes mises à l'échelle pour les graphiques de CPC moyens

Afin de vérifier si les pentes des graphiques de CPC moyens mis à l'échelle convergent vers une valeur constante, il faut simplement répéter le processus précédent avec les données de CPC moyens plutôt que les données de clics. Nous avons obtenu un ensemble de 184 mots-clés qui

fournissent des régressions linéaires acceptables selon nos critères, donc il faut comparer les valeurs de 184 pentes mises à l'échelle. En retirant deux valeurs extrêmes calculées selon le critère de l'étendue interquartile, nous obtenons un total de 182 pentes à comparer. La Figure 5.3 illustre la répartition de ces valeurs de pentes mises à l'échelle.

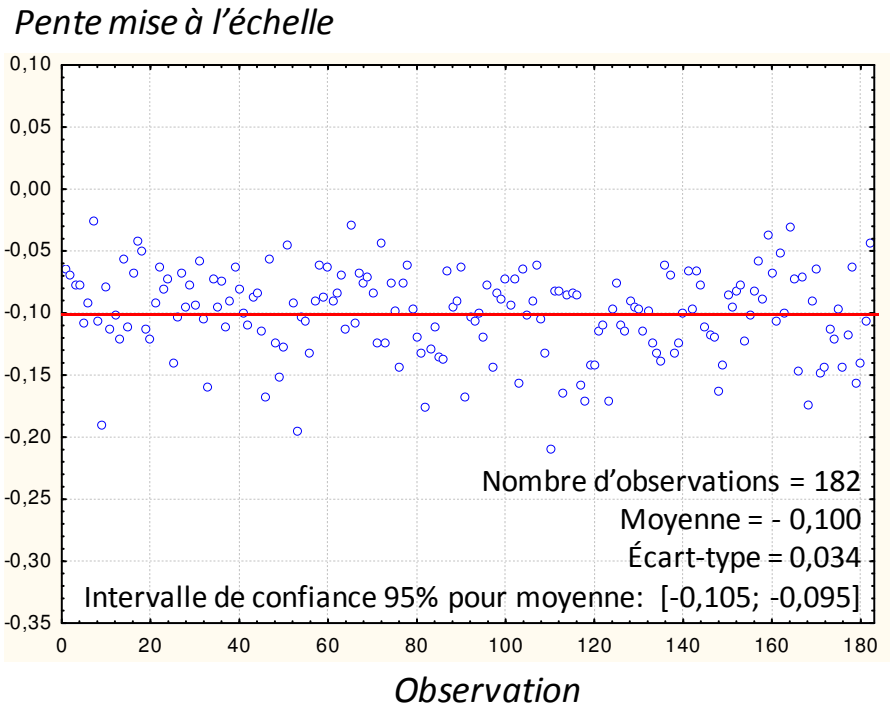



Figure 5.3 : Répartition des pentes mises à l'échelle pour les graphiques de CPC moyen en fonction de la position moyenne

Tout comme dans le cas des clics, il existe une certaine variance autour de la moyenne de -0,100. Cependant, nous considérons que les observations sont suffisamment regroupées pour affirmer que les pentes sont semblables d'un mot-clé à l'autre. L'écart-type de 0,034 démontre que la majorité des observations se retrouvent à l'intérieur d'un intervalle de valeurs relativement restreint.

Prise de décision

Face à ces résultats, nous sommes portés à accepter les deux hypothèses initiales selon lesquelles tous les mots-clés mis à une échelle commune possèdent sensiblement les mêmes taux de décroissance pour leurs fonctions de clics et de CPC moyens. En supposant que tous les mots-clés d'une campagne publicitaire se comporteront comme ceux qui constituent les ensembles à tester, il sera possible de développer des méthodes de prédiction génériques qui utilisent un même taux de décroissance mis à l'échelle pour tous les mots-clés, autant pour la prédiction des clics que pour la prédiction des CPC moyens.

Puisqu'il n'est pas possible d'effectuer des régressions satisfaisantes à partir des mots-clés autres que ceux constituant les ensembles testés, il n'est pas possible de valider une telle méthode en utilisant le principe des régressions. Cependant, nous avons l'intention de vérifier le rendement de cette approche en calculant des taux d'erreur sur l'ensemble des mots-clés des campagnes lorsque nous tentons de prédire leurs comportements à l'aide des taux de décroissance trouvés (-0,158 pour les clics et -0,100 pour les CPC moyens). Ces analyses de marges d'erreurs sont présentées de façon détaillée à la section 5.5.

Application des taux de décroissance génériques aux mots-clés

Étant donné un mot-clé quelconque, le processus pour lui appliquer un taux de décroissance générique est relativement simple. Le raisonnement est le même, autant pour les fonctions de clics que pour les fonctions de CPC. Les valeurs génériques $pente_{clics} = -0{,}158$ et $pente_{cpc} = -0{,}100$ calculées plus tôt seront utilisées pour prédire les nombres de clics et les CPC moyens de ce mot-clé. Puisque ces valeurs représentent les pentes des graphiques lorsque réduits à une échelle commune, il faut ajuster l'échelle de chacun des mots-clés proportionnellement à leur volume.

D'un point de vue mathématique, nous possédons actuellement une valeur de pente que nous cherchons à positionner sur un graphique afin d'obtenir une droite de prédiction. Pour fixer l'emplacement de notre droite, il suffit de déterminer un point par lequel elle passera. Nous choisissons donc de faire passer notre droite par le point moyen de notre graphique. Pour ce faire, nous calculons les valeurs moyennes des positions et des observations sur chacun des graphiques, ce qui nous permet d'obtenir les points moyens $(posMoy, clicsMoy)$ et $(posMoy, cpcMoy)$.

Une fois ce calcul effectué, il suffit de résoudre des systèmes à deux équations et deux inconnues pour obtenir les équations exactes des droites de prédiction. Notons qu'il suffit d'une seule observation pour calculer le point moyen du graphique, donc les fonctions génériques linéaires peuvent s'appliquer à n'importe quel mot-clé possédant au moins une observation dans son historique de données.

Voici les valeurs connues jusqu'à présent :

$penteGenClics$ valeur générique de pente mise à l'échelle calculée selon les analyses initiales, pour la prédiction du nombre de clics;

$penteGenCpc$ valeur générique de pente mise à l'échelle calculée selon les analyses initiales, pour la prédiction des CPC moyens;

$posMoy$ moyenne pondérée des positions moyennes sur chacune des journées $j$

$$posMoy = \frac{\sum_{j=1}^{n} pos_j}{n} ;$$

$clicsMoy$ moyenne des clics sur chacune des journées $j$

$$clicsMoy = \frac{\sum_{j=1}^{n} clics_j}{n} ;$$

$cpcMoy$ moyenne des CPC moyens sur chacune des journées $j$

$$cpcMoy = \frac{\sum_{j=1}^{n} cpc_j}{n} .$$

Voici les valeurs inconnues que nous chercherons à calculer :

| | |
|---|---|
| $clicsPos1$ | nombre de clics prédit en position 1; |
| $cpcPos1$ | CPC moyen prédit en position 1; |
| $pentePredClics$ | pente qui sera appliquée au graphique de clics du mot-clé considéré afin d'obtenir des prédictions en fonction de la position moyenne; |
| $pentePredCpc$ | pente qui sera appliquée au graphique de CPC moyen du mot-clé considéré afin d'obtenir des prédictions en fonction de la position moyenne. |

Nous avons défini la valeur générique de la pente mise à l'échelle comme étant la pente de prédiction d'un mot-clé divisée par sa valeur prédite en position 1. Nous obtenons alors les équations suivantes :

$$penteGenClics = -0{,}158 = \frac{pentePredClics}{clicsPos1} \quad (1)$$

$$penteGenCpc = -0{,}100 = \frac{pentePredCpc}{cpcPos1} \quad (2)$$

Par la suite, sachant que les valeurs prédites décroissent de façon linéaire en fonction de la position (forme $f(x) = m * x + b$ où $m < 0$), nous pouvons déduire les équations suivantes :

$$nombre\ de\ clics\ prédit = pentePredClics * position + nombre\ de\ clics\ prédit\ à\ l'origine$$

$$CPC\ moyen\ prédit = pentePredCpc * position + CPC\ moyen\ prédit\ à\ l'origine$$

En insérant les valeurs des points moyens $(posMoy, clicsMoy)$ et $(posMoy, cpcMoy)$ dans les équations précédentes, nous obtenons :

$$clicsMoy = clicsPos1 + (posMoy - 1) * pentePredClics \quad (3)$$

$$cpcMoy = cpcPos1 + (posMoy - 1) * pentePredCpc \quad (4)$$

Par substitution de (1) dans (3) et (2) dans (4), il est possible de calculer les valeurs des quatre inconnues recherchées.

Exemple d'application de la méthode

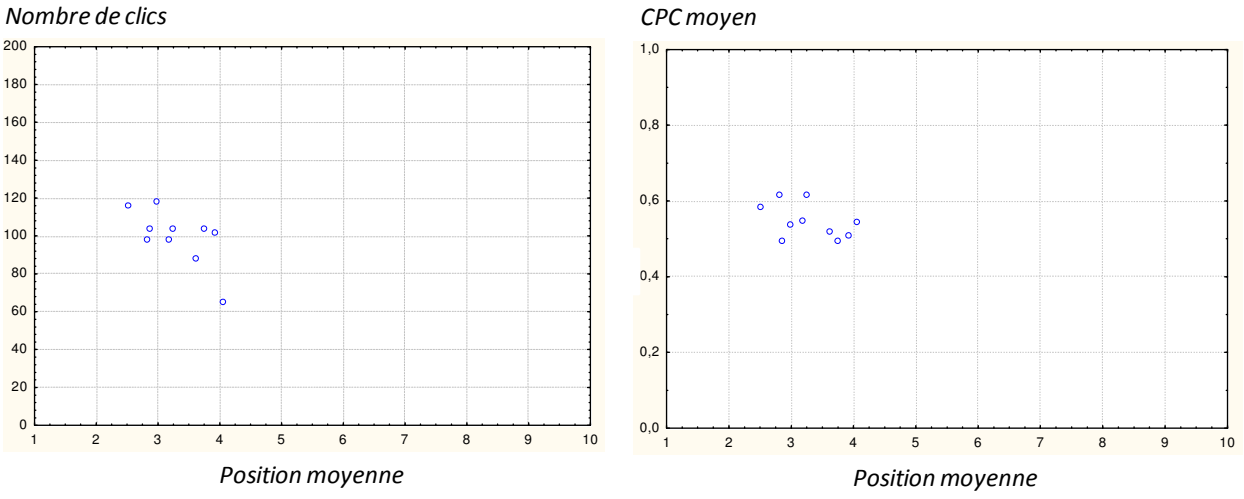
En utilisant les taux de décroissance calculés précédemment, il sera possible de générer des fonctions de prédiction qui estimeront le nombre de clics et les CPC moyens associés à chacune des positions potentielles, pour chacun des mots-clés considérés. Tel qu'expliqué au chapitre 4, les fonctions génériques de prédiction sont destinées à être utilisées pour modéliser le comportement des mots-clés qui ne fournissent pas des régressions adéquates, qui ne possèdent pas un nombre d'observations récentes ou qui ne totalisent pas un nombre de clics suffisant (voir algorithme de classification à la Figure 4.5). Nous présentons donc un exemple qui illustre comment les fonctions génériques seront appliquées à ce type de mot-clé.

Considérons un mot-clé qui possède un historique de 10 journées de données, toutes avec des positions moyennes relativement stables. Les données de ce mot-clé sont représentées au Tableau 5.3.

Tableau 5.3 : Données historiques de position moyenne, nombre de clics et CPC moyen d'un mot-clé quelconque

| jour | position moyenne | nombre de clics | CPC moyen |
|---|---|---|---|
| 1 | 3,74 | 104 | 0,49 |
| 2 | 2,98 | 118 | 0,54 |
| 3 | 2,52 | 116 | 0,59 |
| 4 | 2,86 | 104 | 0,49 |
| 5 | 2,82 | 98 | 0,62 |
| 6 | 3,24 | 104 | 0,62 |
| 7 | 3,93 | 102 | 0,51 |
| 8 | 3,17 | 98 | 0,55 |
| 9 | 3,62 | 88 | 0,52 |
| 10 | 4,06 | 65 | 0,54 |
| **moyennes** | **3,29** | **99,70** | **0,55** |

Pour inclure ce mot-clé dans le modèle d'optimisation, il est nécessaire d'obtenir des prédictions de clics et de CPC moyen pour chacune des positions auxquelles il peut se retrouver. La Figure 5.4 illustre graphiquement la répartition des 10 observations actuellement disponibles. Visiblement, il n'est pas possible de générer des régressions satisfaisantes à partir de ces données. Les fonctions génériques de prédiction semblent donc représenter une alternative intéressante pour prédire le comportement de ce mot-clé.




(A) (B)

Figure 5.4 : Représentation graphique des données historiques de clics (A) et de CPC moyen (B) en fonction de la position moyenne pour un mot-clé quelconque

Voici les valeurs connues jusqu'à présent :

| | |
|---|---|
| $penteGenClics$ | -0,158 |
| $penteGenCpc$ | -0,100 |
| $posMoy$ | 3,29 (voir Tableau 5.3) |
| $clicsMoy$ | 99,70 (voir Tableau 5.3) |
| $cpcMoy$ | 0,55 (voir Tableau 5.3) |

Insérons ces valeurs dans les équations 1 à 4 :

$$-0{,}158 = \frac{pentePredClics}{clicsPos1} \quad (1)$$

$$-0{,}100 = \frac{pentePredCpc}{cpcPos1} \quad (2)$$

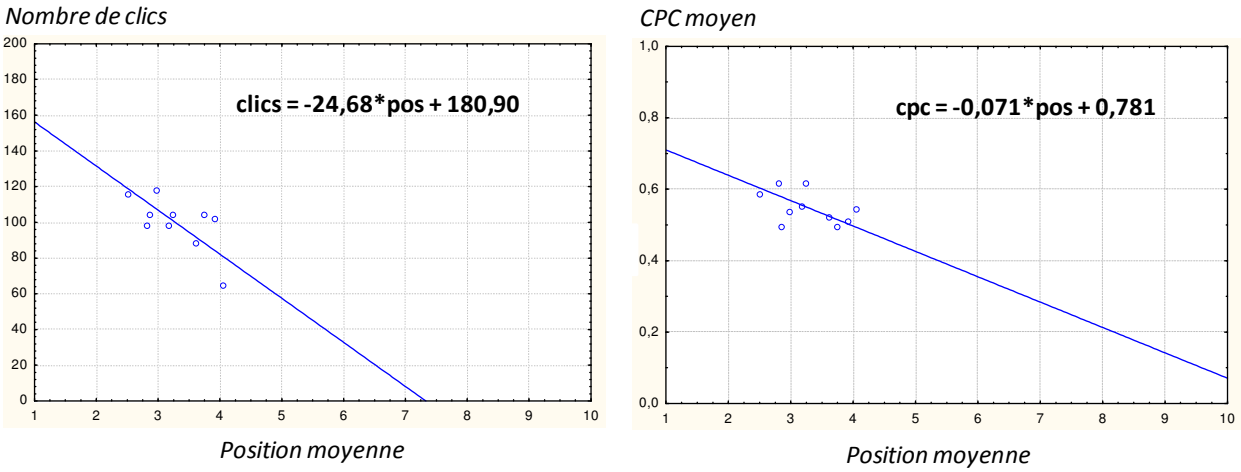
$$99{,}70 = clicsPos1 + (3{,}29 - 1) * pentePredClics \quad (3)$$

$$0{,}55 = cpcPos1 + (3{,}29 - 1) * pentePredCpc \quad (4)$$

Par substitution de (1) dans (3) et (2) dans (4), il est possible de calculer les valeurs suivantes :

| | |
|---|---|
| $clicsPos1$ | 156,22 clics |
| $cpcPos1$ | 0,71 $ |
| $pentePredClics$ | -24,68 clics/position |
| $pentePredCpc$ | -0,071 $/position |

La Figure 5.5 montre les fonctions de prédiction obtenues à partir de ces valeurs. Ces fonctions permettent d'estimer, malgré l'insuffisance de données historiques, le comportement du mot-clé en question.




(A) (B)

Figure 5.5 : Représentation graphique des fonctions génériques linéaires pour la prédiction des clics (A) et du CPC moyen (B) en fonction de la position moyenne pour un mot-clé quelconque

Ajustement des fonctions de prédiction

Avec les valeurs génériques $penteGenClics = -0{,}158$ et $penteGenCpc = -0{,}100$ obtenues lors des analyses initiales, il n'est pas possible d'estimer toutes les positions possibles. En effet, la valeur de la pente générique de clics fait en sorte que la fonction de prédiction des clics prédit des valeurs négatives à partir de la position 7,33 :

$$abscisse\ \text{à}\ l'origine = \frac{ordonnée\ \text{à}\ l'origine}{-pente} = \frac{1 + 0{,}158}{0{,}158} = 7{,}33$$

De plus, la valeur de la pente générique de CPC moyen fait en sorte que la fonction de prédiction des CPC moyens prédit des valeurs négatives à partir de la position 11 :

$$abscisse\ \text{à}\ l'origine = \frac{ordonnée\ \text{à}\ l'origine}{-pente} = \frac{1 + 0{,}100}{0{,}100} = 11$$

Sachant cela, nous devons ajuster nos fonctions de prédiction avant de les inclure dans le modèle d'optimisation, afin de s'assurer de fournir des données initiales cohérentes. Puisque notre modèle se limite aux positions 1 à 10, la prédiction de valeurs négatives dans les positions supérieures à 11 pour la fonction de CPC moyen ne pose pas de problème. Cependant, pour ajuster la fonction de clics, nous choisissons de prédire des valeurs nulles pour toutes les positions plus grandes que le seuil calculé (7,33). Cela risque d'engendrer certaines erreurs de prédiction, mais ces erreurs risquent d'être relativement peu importantes considérant que les nombres de clics sont généralement très faibles dans les positions basses.

Nous devons également gérer les cas rares où la position moyenne calculée à partir des observations excède le seuil de position critique (lorsque $posMoy > 7{,}33$ pour les clics ou lorsque $posMoy > 11$ pour les CPC). Dans ces cas, les pentes de prédiction $pentePredClics$ et $pentePredCpc$ calculées selon les équations mentionnées plus tôt sont toujours positives. Cela ne respecte pas nos hypothèses initiales selon lesquelles les clics et les CPC moyens sont décroissants en fonction de la position moyenne. Nous choisissons donc de prédire le comportement de ces mots-clés avec une fonction constante qui ne dépend pas de la position : $clics\ prédits = clicsMoy$ ou $CPC\ prédit = cpcMoy$, selon le cas.

Calcul de la position moyenne

Dans le cas des mots-clés qui possèdent une ou plusieurs journées où le nombre de clics observé est nul, il faut calculer des valeurs de position moyenne ($posMoy$) différentes pour la fonction de clics et la fonction de CPC moyen. En effet, il est possible de tracer une observation où le nombre de clics est nul sur un graphique de clics, mais cela n'est pas possible sur un graphique de CPC moyen; le calcul du CPC moyen engendre une division par zéro lorsque le nombre de clics est nul ($\text{CPC} = \text{coût quotidien}/\text{nombre de clics}$). Bref, si chaque journée de l'historique possède au moins un clic (comme dans l'exemple précédent), nous avons :

$$posMoyClics = posMoyCpc = posMoy$$

Sinon, nous devons considérer deux valeurs de position moyenne différentes qui ne sont pas nécessairement égales : $posMoyClics$ et $posMoyCpc$.

Exemples de comparaison entre les régressions et les fonctions génériques

Afin de comparer les méthodes de prédiction génériques avec les prédictions calculées par régression, nous avons choisi de représenter graphiquement les fonctions obtenues avec chacune de ces deux méthodes, pour chacun des mots-clés faisant partie des ensembles à tester (204 mots-clés pour la prédiction des clics et 182 mots-clés pour la prédiction des CPC moyens). Quelques exemples de graphiques choisis au hasard sont présentés à l'Annexe 4. Dans certains cas, les valeurs estimées par la fonction générique sont très différentes de celles prédites par la fonction de régression. Cependant, dans la majorité des cas, ces deux fonctions sont relativement semblables et nous considérons que les résultats obtenus sont satisfaisants dans l'ensemble. Ils seront analysés de façon quantitative à la section 5.5.

## 5.4 Fonctions génériques exponentielles

Même si la méthode linéaire semble fournir des résultats satisfaisants, nous avons choisi de tester une seconde méthode qui suppose des taux de décroissance exponentiels plutôt que linéaires. Cette approche est très semblable à la précédente, car il suffit de linéariser l'équation fournie par la régression exponentielle pour obtenir une équation sous la forme linéaire. Bref, l'analyse de cette section est structurée de la même façon que celle de la section 5.3, mais nous utilisons des régressions exponentielles plutôt que linéaires pour modéliser les fonctions de clics et de CPC moyen des mots-clés.

Linéarisation

En effectuant une régression exponentielle, les équations obtenues sont de la forme suivante :

$$clics = k_{clics} * e^{m*pos} \quad (5)$$

$$cpc = k_{cpc} * e^{n*pos} \quad (6)$$

où $k_{clics}$ et $k_{cpc}$ sont les valeurs à l'origine de chacune des fonctions, $m$ et $n$ sont des paramètres qui caractérisent le rythme de décroissance des fonctions.

Pour linéariser ces équations, il suffit de prendre le logarithme de chaque côté des égalités :

$$\ln(clics) = \ln(k_{clics} * e^{m*pos})$$

$$\ln(cpc) = \ln(k_{cpc} * e^{n*pos})$$

Après quelques simplifications, nous obtenons les équations suivantes :

$$\ln(clics) = m * pos + \ln(k_{clics}) \quad (7)$$

$$\ln(cpc) = n * pos + \ln(k_{cpc}) \quad (8)$$

C'est cette dernière forme d'équation qui est utilisée pour calculer les régressions à partir des données. Plus précisément, nous effectuons des régressions linéaires avec les observations $(pos_j, \ln(clics_j))$ et $(pos_j, \ln(cpc_j))$ pour chacune des journées $j$ considérées dans la période historique.

Ajustement d'échelle

Contrairement au cas linéaire, il n'est pas nécessaire de mettre tous les graphiques sur une échelle commune afin de comparer leurs taux de décroissance. En effet, dans les équations exponentielles (5) et (6), les paramètres $m$ et $n$ caractérisent la décroissance relative des valeurs d'une position à l'autre et sont indépendantes de l'échelle utilisée. Ce sont plutôt les valeurs initiales $k_{clics}$ et $k_{cpc}$ qui déterminent l'échelle de grandeur pour chaque graphique. Il suffit donc de comparer les valeurs des paramètres $m$ et $n$ de chacune des régressions calculées pour déterminer si les taux de décroissance des fonctions sont sensiblement les mêmes d'un mot-clé à l'autre.

Cependant, afin de rendre le problème plus intuitif, nous utilisons une méthode équivalente pour comparer les taux de décroissance des fonctions. Appelons $c_{clics} = e^m$ et $c_{cpc} = e^n$. Les valeurs $c_{clics}$ et $c_{cpc}$ représentent la proportion de clics et de CPC moyen qui est conservée d'une position à l'autre, lorsqu'on se déplace d'une position vers le bas dans la liste. Par exemple, une valeur de $m = -0{,}40$ nous donnerait $c_{clics} = e^{-0{,}40} = 0{,}67$, ce qui impliquerait que le nombre de clics du mot-clé en question décroît en moyenne de 33% lorsqu'il se déplace d'une position vers le bas. Considérant que les nombre de clics et CPC moyens sont décroissants en fonction de la position, les valeurs de $c_{clics}$ et $c_{cpc}$ devraient se retrouver entre 0 et 1. Des valeurs se

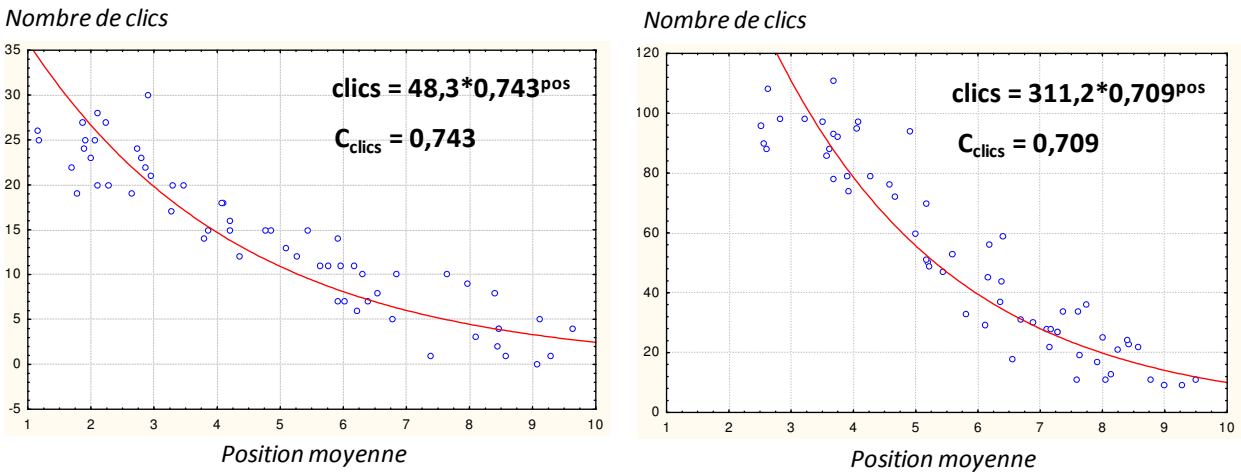
rapprochant de 0 impliqueront une décroissance rapide et des valeurs plus près de 1 représenteront une décroissance plus lente.

En utilisant ces substitutions, nous cherchons donc à déterminer s'il existe des valeurs génériques de $c_{clics}$ et $c_{cpc}$ qui pourraient caractériser la décroissance des mots-clés, à partir des formules suivantes :

$$clics = k_{clics} * (c_{clics})^{pos} \quad (9)$$

$$cpc = k_{cpc} * (c_{cpc})^{pos} \quad (10)$$

La Figure 5.6 montre l'application de ce principe aux deux mêmes mots-clés qui ont été utilisés pour illustrer la méthode linéaire à la Figure 5.1. En utilisant des régressions exponentielles, il est possible de constater que les coefficients de décroissance $c_{clics}$ des deux mots-clés sont très semblables, soit de 0,743 et 0,709.




(A) (B)

Figure 5.6 : Exemple de graphiques de clics pour deux mots-clés (A et B) avec taux de décroissance relatifs semblables

Extraction d'ensembles de mots-clés à tester

Les ensembles de mots-clés à tester sont obtenus de la même façon qu'à la section précédente. Les critères utilisés sont tous les mêmes, à l'exception du critère de coefficient de détermination. Plutôt que d'imposer un $R^2$ de régression linéaire $\geq 0{,}30$, nous exigeons un $R^2$ de régression exponentielle linéarisée $\geq 0{,}30$. Rappelons que la régression exponentielle linéarisée est obtenue en effectuant une régression linéaire sur les valeurs de $\ln(nombre\ de\ clics)$ et $\ln(CPC\ moyen)$ en fonction de la position moyenne. En utilisant ces critères, nous réussissons à extraire un ensemble de 211 mots-clés pour l'analyse des clics et 185 mots-clés pour l'analyse des CPC moyens.

Analyse des coefficients de décroissance pour les graphiques de clics

Nous cherchons à déterminer si les valeurs de $c_{clics}$ sont semblables d'un mot-clé à l'autre. Pour y arriver, il suffit d'analyser la répartition de ces valeurs pour chacun des mots-clés qui constituent l'ensemble à tester.

La Figure 5.7 illustre la répartition des 211 valeurs de $c_{clics}$ obtenues. Il n'existe aucune valeur extrême et la moyenne des observations est de 0,613. Considérant l'écart-type de 0,103 qui est relativement faible, nous jugeons qu'il serait acceptable d'utiliser la valeur moyenne pour approximer le taux de décroissance des clics de l'ensemble des mots-clés. En utilisant une telle valeur générique, nous estimons qu'en moyenne 61,3% des clics sont conservés lorsqu'un mot-clé descend d'une position, peu importe la position.

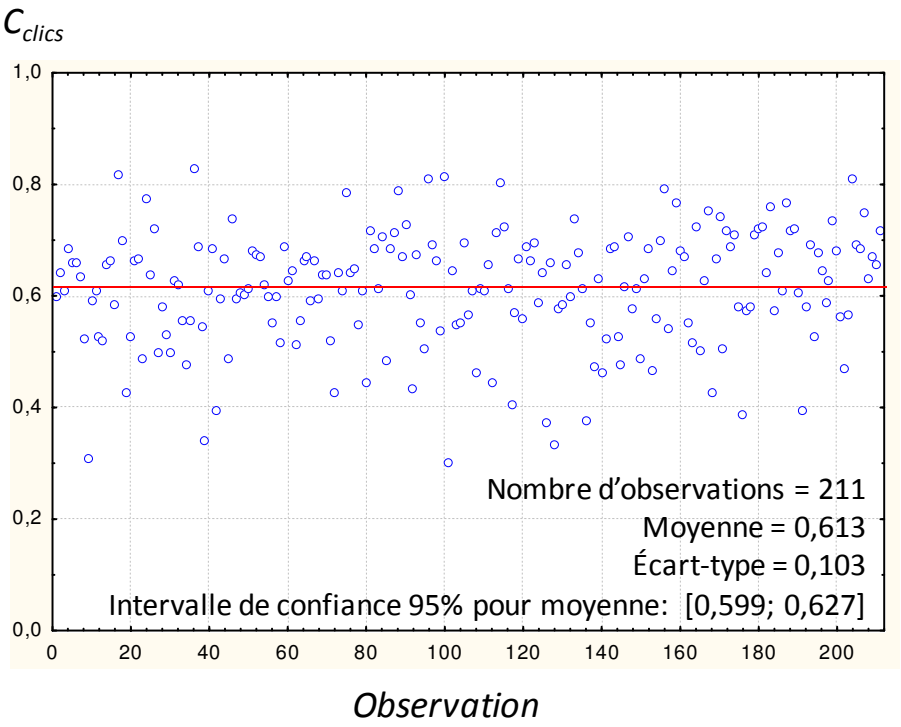



Figure 5.7 : Répartition des coefficients de décroissance $c_{clics}$ pour les graphiques de nombre de clics en fonction de la position moyenne

Analyse des coefficients de décroissance pour les graphiques de CPC moyen

Il faut déterminer si les valeurs de $c_{cpc}$ sont relativement constantes d'un mot-clé à l'autre. Tout comme avec les fonctions de clics, nous analysons la répartition des valeurs pour chacun des mots-clés de l'ensemble à tester.

En utilisant le critère de l'étendue interquartile, nous constatons qu'il existe trois valeurs extrêmes parmi nos 185 observations de $c_{cpc}$ obtenues. Après avoir enlevé ces trois valeurs extrêmes, nous observons la répartition illustrée à la Figure 5.8. La moyenne des observations est de 0,843. Comparativement aux autres résultats présentés pour les $c_{clics}$, nous constatons que la répartition des $c_{cpc}$ est encore plus serrée autour de la moyenne. L'écart-type qui caractérise la dispersion des données est seulement de 0,068.

Compte tenu de ces résultats, nous jugeons qu'il serait acceptable d'utiliser la valeur moyenne pour approximer le taux de décroissance du CPC moyen de l'ensemble des mots-clés. En utilisant

une telle valeur générique, nous estimons qu'en moyenne 84,3% des CPC moyens sont conservés lorsqu'un mot-clé descend d'une position, peu importe la position.

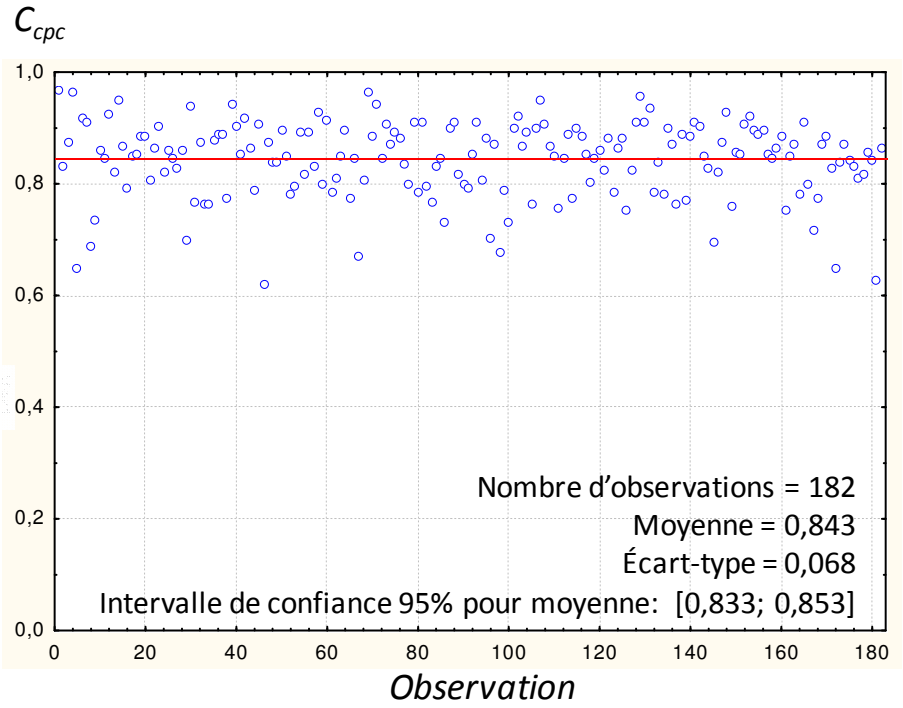



Figure 5.8 : Répartition des coefficients de décroissance $c_{cpc}$ pour les graphiques de CPC moyen en fonction de la position moyenne

Prise de décision

En considérant les résultats obtenus, nous jugeons qu'il est acceptable d'approximer les coefficients de décroissance des mots-clés en utilisant les valeurs moyennes de $c_{clics} = 0{,}613$ et $c_{cpc} = 0{,}843$ calculées plus haut. Cette approximation risque de générer certaines erreurs, particulièrement dans le cas des clics, car les observations ne sont pas toujours parfaitement resserrées autour de la moyenne. Cependant, les méthodes de prédiction génériques sont destinées à être utilisées en l'absence de fonctions de régression satisfaisantes. Dans ces cas, nous nous contenterons de prédire avec les fonctions génériques plutôt que de ne rien prédire du tout.

La performance de l'approche générique exponentielle sera analysée plus en détail à la section 5.5. Dans cette section, nous déterminerons de façon quantitative si l'utilisation des coefficients de décroissance calculés (0,613 pour les clics et 0,843 pour les CPC moyens) fournit des résultats

acceptables pour l'ensemble des mots-clés d'une campagne publicitaire. Cela nous permettra de valider les décisions qui viennent d'être prises concernant la convergence des valeurs de $c_{clics}$ et $c_{cpc}$.

Application des taux de décroissance génériques aux mots-clés

Pour appliquer les taux de décroissance génériques à un mot-clé quelconque, il suffit de posséder au moins une observation récente dans l'historique de données de ce mot-clé. Tout comme avec la méthode linéaire, nous cherchons à appliquer des taux de décroissance fixes à un mot-clé donné :

$$c_{clics} = 0{,}613$$

$$c_{cpc} = 0{,}843$$

Il faut donc ajuster l'échelle des données auxquelles cette décroissance sera appliquée. Dans les équations (9) et (10) mentionnées précédemment, cette échelle est déterminée par les valeurs $k_{clics}$ et $k_{cpc}$, qui sont des constantes propres à chaque mot-clé (elles représentent les valeurs à l'origine pour les fonctions de clics et de CPC moyen). Afin de calculer ces valeurs pour chacun des mots-clés, nous devons utiliser $(posMoy, clicsMoy)$ et $(posMoy, cpcMoy)$, les points moyens des observations :

$$posMoy = \frac{\sum_{j=1}^{n} pos_j}{n}$$

$$clicsMoy = \frac{\sum_{j=1}^{n} clics_j}{n}$$

$$cpcMoy = \frac{\sum_{j=1}^{n} cpc_j}{n}$$

Une fois ces valeurs moyennes calculées, il suffit d'isoler les valeurs de $k_{clics}$ et $k_{cpc}$ :

$$k_{clics} = \frac{clicsMoy}{(c_{clics})^{posMoy}}$$

$$k_{cpc} = \frac{cpcMoy}{(c_{cpc})^{posMoy}}$$

Nous obtenons alors les équations des fonctions de prédiction sous la même forme que les équations (9) et (10) :

$$clics = k_{clics} * (c_{clics})^{pos}$$
$$cpc = k_{cpc} * (c_{cpc})^{pos}$$

où $k_{clics}$ et $k_{cpc}$ sont des constantes propres à chaque mot-clé, $c_{clics}$ et $c_{cpc}$ sont des constantes globales, $clics$ et $cpc$ sont les variables dépendantes et $pos$ est la variable indépendante.

Exemple d'application de la méthode

Afin d'illustrer comment ce type de fonction peut être utilisé pour prédire le comportement des clics et des CPC moyens en fonction de la position, nous présentons un exemple à partir du même mot-clé utilisé pour l'exemple de la section précédente. Les données de ce mot-clé sont présentées au Tableau 5.3 et une représentation graphique de ses données historiques est fournie à la Figure 5.4.

Voici les valeurs connues jusqu'à présent :

| | |
|---|---|
| $c_{clics}$ | 0,613 |
| $c_{cpc}$ | 0,843 |
| $posMoy$ | 3,29 (voir Tableau 5.3) |
| $clicsMoy$ | 99,70 (voir Tableau 5.3) |
| $cpcMoy$ | 0,55 (voir Tableau 5.3) |

Par la suite, il suffit d'isoler les valeurs de $k_{clics}$ et $k_{cpc}$ :

$$k_{clics} = \frac{clicsMoy}{(c_{clics})^{posMoy}} = \frac{99{,}70}{(0{,}613)^{3{,}29}} = 498{,}82\ clics$$

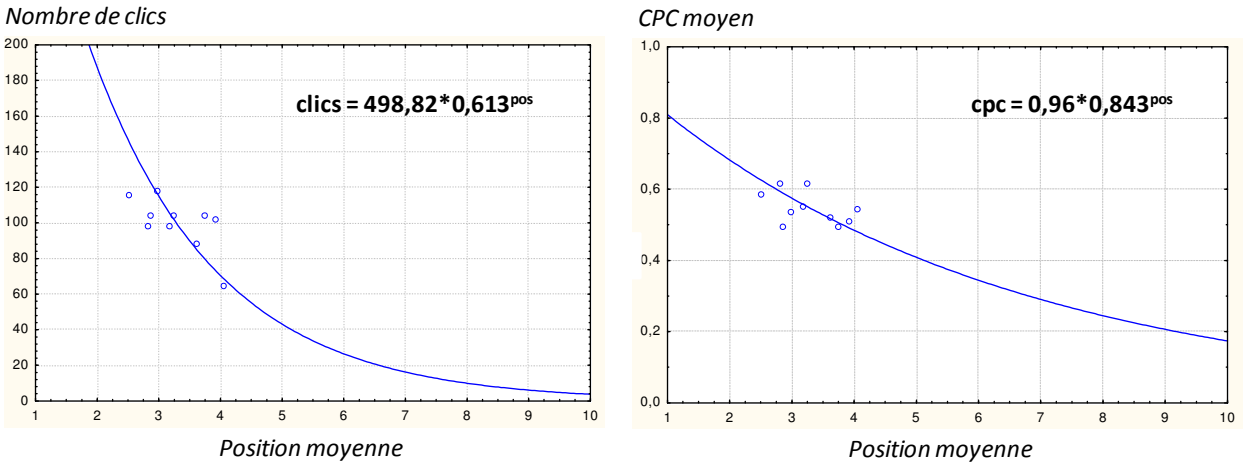
$$k_{cpc} = \frac{cpcMoy}{(c_{cpc})^{posMoy}} = \frac{0{,}55}{(0{,}843)^{3{,}29}} = 0{,}96\ \$/clic$$

Nous obtenons donc les équations génériques de prédiction suivantes, pour ce mot-clé :

$$clics = 498{,}82 * 0{,}613^{pos}$$

$$cpc = 0{,}96 * 0{,}843^{pos}$$

Ces fonctions sont représentées graphiquement à la Figure 5.9. Malgré le manque de données historiques, les fonctions génériques nous permettent de prédire des valeurs de clics et de CPC moyens pour chacune des positions potentielles.




(A) (B)

Figure 5.9 : Représentation graphique des fonctions génériques linéaires pour la prédiction des clics (A) et du CPC moyen (B) en fonction de la position moyenne pour un mot-clé quelconque

Calcul de la position moyenne

Tout comme dans le cas linéaire, il faut considérer deux valeurs de position moyenne différentes ($posMoyClics$ et $posMoyCpc$) si au moins une des journées de la période historique d'un mot-clé possède un nombre de clics nul. En effet, dans un tel cas, la position moyenne du graphique de clics n'est pas nécessairement la même que celle du graphique de CPC moyen.

Plage de positions considérées

En effectuant nos analyses, nous avons constaté que les fonctions de prédiction obtenues prédisaient des valeurs de clics et de CPC moyen beaucoup trop élevées lorsque la valeur de la position moyenne ($posMoy$) utilisée pour centrer la courbe était trop grande. Puisque nous supposons des taux de variation exponentiels, il faut éviter d'utiliser des valeurs de positions moyennes trop élevées lors du calcul de $k_{clics}$ et $k_{cpc}$ afin de ne pas obtenir des prédictions exagérées. Suite à plusieurs séries de tests, nous avons constaté que la méthode performe mieux lorsque nous traitons différemment les mots-clés où $posMoy > 10$. Dans ces rares cas, nous nous contentons de prédire des valeurs de clics et de CPC moyen constantes, en utilisant la moyenne des observations disponibles. Cela nous permet de prédire des valeurs pour la totalité des mots-clés des campagnes publicitaires, ce qui est nécessaire pour effectuer une comparaison équitable entre la méthode exponentielle et la méthode linéaire (cette comparaison est présentée à la section suivante).

Exemples de comparaison entre les régressions et les fonctions génériques

Tout comme avec le cas linéaire, nous voulions comparer graphiquement les fonctions obtenues par les fonctions génériques avec celles obtenues par régression. À l'Annexe 5, nous présentons quelques exemples de graphiques choisis au hasard, parmi les ensembles initiaux de mots-clés (211 mots-clés pour les clics et 182 pour les CPC moyens). De façon générale, nous sommes satisfaits de ces résultats. Nous analysons plus en détail la performance des fonctions génériques sur l'ensemble des mots-clés d'une campagne publicitaire à la section 5.5.

## 5.5 Comparaison des deux approches

Critères de comparaison

Afin d'analyser la performance des méthodes de prédiction génériques linéaire et exponentielle, nous avons choisi de comparer la qualité de leurs prédictions sur une période de 30 jours consécutifs (période de temps distincte de la période historique utilisée pour construire les fonctions). Pour obtenir une évaluation juste et objective de chacune des méthodes, il est nécessaire de développer divers critères quantifiables qui se calculent à partir des valeurs prédites et observées de chacune des journées de comparaison.

D'abord, en vue de maximiser le rendement de notre modèle d'optimisation, il est important de minimiser l'erreur associée à la prédiction du nombre total de clics et des coûts totaux quotidiens. En effet, l'application des solutions obtenues par le modèle risque de fournir des résultats plus satisfaisants si les prédictions quotidiennes sont justes. C'est pourquoi nous avons choisi d'utiliser des critères qui quantifient l'erreur moyenne associée à chacune de ces prédictions :

1) Erreur absolue moyenne associée à la prédiction des clics $= \frac{\sum_{j=1}^{30} \left| clicsPred_j - clicsObs_j \right|}{\sum_{j=1}^{30} clicsObs_j}$,

2) Erreur absolue moyenne associée à la prédiction des coûts $= \frac{\sum_{j=1}^{30} \left| co\hat{u}tsPred_j - co\hat{u}tsObs_j \right|}{\sum_{j=1}^{30} co\hat{u}tsObs_j}$,

où $j$ est l'indice indiquant la journée prédite ou observée ($j$ = 1 à 30), $clicsPred_j$ et $co\hat{u}tsPred_j$ sont les valeurs de clics et de coûts prédites à la journée $j$, $clicsObs_j$ et $co\hat{u}tsObs_j$ sont les valeurs de clics et coût observées à la journée $j$.

*Remarque #1 :* Afin que ces deux critères évaluent les prédictions de clics et CPC moyen de façon indépendante, les coûts prédits du critère 2 sont calculés à partir des clics observés et non à partir des clics prédits. Plus spécifiquement, $co\hat{u}tsPred_j = clicsObs_j * cpcPred_j$. De cette façon, le critère 2 fournit uniquement une estimation de la qualité de prédiction des fonctions de CPC moyen. Un critère d'erreur qui calculerait les coûts prédits à partir des clics prédits tiendrait compte à la fois des erreurs de prédiction des clics et des erreurs de prédiction des CPC moyens dans le même critère.

*Remarque #2 :* Nous aurions pu utiliser un critère qui calcule l'erreur relative sur les CPC moyens plutôt que sur les coûts quotidiens, mais considérant la structure du modèle d'optimisation, nous croyons qu'il est plus intéressant d'évaluer les prédictions de coûts sur une base quotidienne. En effet, afin d'évaluer la profitabilité de chacune des positions, le modèle utilise ultimement des prédictions quotidiennes pour les clics totaux et les coûts totaux.

*Remarque #3 :* Puisque nous divisons la somme des erreurs absolues par la somme des valeurs observées, l'erreur calculée exprime l'erreur moyenne qui sera associée à chacune des prédictions. Par exemple, une valeur d'erreur absolue de 0,30 implique que les prédictions se trompent, en moyenne, de 30% par rapport aux observations.

Suite à la définition de ces deux premiers critères, nous cherchions également à prendre en considération l'exactitude des prédictions sur une période prolongée plutôt que sur une base individuelle. Du point de vue des annonceurs, il est important que les budgets prévus soient respectés et que le nombre total de clics estimé soit relativement fiable. Cela étant dit, les deux critères d'erreur précédents sont calculés de façon absolue, donc ils ne permettent pas de détecter si une méthode surestime ou sous-estime constamment le nombre de clics ou les coûts. C'est pourquoi nous ajoutons deux autres critères qui quantifient l'effet de surestimation ou de sous-estimation :

$$3)\ \text{Erreur moyenne associée à la prédiction des clics} = \frac{\sum_{j=1}^{30}(clicsPred_j - clicsObs_j)}{\sum_{j=1}^{30} clicsObs_j},$$

$$4)\ \text{Erreur moyenne associée à la prédiction des coûts} = \frac{\sum_{j=1}^{30}(coûtsPred_j - coûtsObs_j)}{\sum_{j=1}^{30} coûtsObs_j},$$

où $j$ est l'indice indiquant la journée prédite ou observée ($j$=1 à 30), $clicsPred_j$ et $coûtsPred_j$ sont les valeurs de clics et de coûts prédites à la journée $j$, $clicsObs_j$ et $coûtsObs_j$ sont les valeurs de clics et coût observées à la journée $j$.

*Remarque #1 :* Tout comme avec le critère 2, les coûts prédits utilisés pour le critère 4 sont calculés à partir des clics observés et non à partir des clics prédits, afin que les critères 3 et 4 évaluent les fonctions de clics et de CPC moyen indépendamment.

*Remarque #2 :* Puisque nous divisons la somme des erreurs par la somme des valeurs observées, l'erreur calculée exprime le pourcentage de surestimation ou de sous-estimation global. Par exemple, une valeur d'erreur totale de -0,30 implique que nous prédisons, en moyenne, des valeurs 30% plus faibles que les valeurs réellement observées.

Comparaison quantitative des méthodes

Afin de s'assurer que l'application des fonctions génériques de prédiction peut s'étendre à l'ensemble des mots-clés qui constituent une campagne publicitaire, nous avons choisi de calculer ces quatre critères d'erreur sur chacune des 20 banques de données présentées à la section 5.2. De cette façon, nous pouvons obtenir une estimation des erreurs générées par les prédictions. De plus, nous pouvons utiliser les résultats obtenus pour comparer les deux méthodes de prédiction (linéaire et exponentielle) et choisir celle qui fonctionne le mieux, autant pour la prédiction des clics que pour la prédiction des CPC moyens.

Le Tableau 5.4 montre les résultats globaux agrégés obtenus suite à l'application des méthodes de prédiction linéaire et exponentielle aux fonctions de clics et de CPC moyen, pour l'ensemble des 20 banques de données disponibles. Puisque les critères d'évaluation sont indépendants, nous pouvons évaluer les fonctions de clics et de CPC moyen séparément. Les résultats démontrent, malgré des performances très semblables, que la méthode exponentielle performe mieux que la méthode linéaire dans tous les cas. L'analyse quantitative des méthodes révèle donc que les fonctions génériques exponentielles sont plus performantes que les fonctions génériques linéaires, selon les quatre critères d'évaluation que nous avons développés.

Tableau 5.4 : Résultats globaux agrégés sur 20 banques de données pour les critères d'évaluation des fonctions de clics et de CPC moyen

| fonction de clics | critère # 1 | critère # 3 | fonction de CPC moyen | critère # 2 | critère # 4 |
|---|---|---|---|---|---|
| linéaire | 0,5340 | -0,0440 | linéaire | 0,3080 | -0,1180 |
| exponentielle | 0,5270 | -0,0346 | exponentielle | 0,3066 | -0,1058 |

Évidemment, les marges d'erreur indiquées sont relativement élevées, particulièrement pour les critères 1 et 2. Avec ces résultats, nous estimerions que notre meilleure méthode de prédiction se trompe en moyenne de 52,70% à chaque prédiction de clics et de 30,66% à chaque prédiction de coût. Nous considérons que de telles marges d'erreur sont inacceptables et ne pensons pas que le modèle d'optimisation puisse améliorer le rendement des campagnes publicitaires si les prédictions obtenues sont aussi imprécises. Cependant, ces marges d'erreur sont calculées uniquement dans le but de comparer les méthodes de prédiction linéaire et exponentielle entre elles. Une fois que la meilleure méthode sera identifiée, nous pourrons lui appliquer divers raffinements dans le but de réduire les marges d'erreur globales associées à chacun des quatre critères d'évaluation. Ces raffinements sont expliqués à la section 5.6. De plus, il est important de mentionner que certains ajustements ont été apportés à chacune des méthodes afin de leur permettre de prédire des valeurs pour la totalité des mots-clés d'une campagne. Cela visait à fournir une comparaison équitable entre les deux méthodes, mais il est évident que ces ajustements pourront être modifiés afin de réduire les taux d'erreur.

Comparaison qualitative des méthodes

Certains avantages et inconvénients associés aux méthodes ne peuvent être pris en compte par des critères quantitatifs. Il est toutefois important de considérer ces facteurs afin d'effectuer un choix de façon intelligente.

D'abord, la méthode linéaire ne fournit pas des prédictions pour toutes les positions. Puisque les droites de prédiction de clics atteignent 0 en position 7,33, nous devons prédire des valeurs de clics nulles pour toutes les positions subséquentes. Cela équivaut à ignorer ces positions, car le modèle d'optimisation ne choisira pas des positions où le nombre de clics prédit est nul. De plus, la méthode linéaire démontre un comportement instable autour de ce seuil de position critique. Lorsque la position moyenne de l'historique est plus grande que 7,33, la pente de prédiction obtenue est positive. Du point de vue de l'optimisation, il est inacceptable d'avoir des fonctions de prédiction de clics ou de CPC moyen croissantes. Nous devons donc prendre des mesures particulières afin d'éviter de tels cas.

D'un autre côté, la méthode exponentielle implique de traiter comme exception les mots-clés rares où la position moyenne des observations se trouve dans des valeurs trop élevées.

Cependant, les fonctions exponentielles offrent l'avantage de prédire des valeurs non nulles et décroissantes sur tout leur domaine. Elles permettent donc d'effectuer des prédictions pour chacune des positions considérées (1 à 10). Elles sont relativement faciles à appliquer et nous considérons que leur fonctionnement est très intuitif et simple à expliquer; le coefficient de décroissance représente concrètement la proportion de clics ou de CPC moyen qui sera conservée lorsque l'annonce du mot-clé glissera d'une position vers le bas.

Finalement, tel que mentionné plus tôt, il est important de souligner que les références qui traitent de ce sujet dans la littérature utilisent des fonctions exponentielles pour prédire le comportement de leurs mots-clés. Autant pour les clics que pour les CPC moyens, les études effectuées jusqu'à présent ont recours à des hypothèses de décroissance exponentielle pour modéliser le comportement de chacun de leurs mots-clés (Kitts & Leblanc, 2004; Kitts et al., 2005; Ganchev et al., 2007). Bref, en se fiant aux études publiées concernant l'estimation des fonctions de clics et de CPC moyen, il est clair que les fonctions exponentielles semblent être les plus communément utilisées.

Choix de la meilleure méthode

Suite à l'analyse des méthodes en fonction de critères quantitatifs ainsi qu'en fonction de facteurs non quantifiables, il apparaît évident que la méthode exponentielle représente le meilleur choix. En effet, elle performe mieux que la méthode linéaire selon chacun des quatre critères d'évaluation que nous avons définis. De plus, elle limite les cas d'exceptions et offre des fonctions de prédiction qui sont simples et faciles à appliquer. Finalement, la fonction exponentielle semble être le type de fonction préféré par les chercheurs dans la littérature pour prédire les variables de clics et de CPC moyen en fonction de la position. En considérant tous ces facteurs, nous choisissons de n'utiliser que les fonctions génériques exponentielles pour la suite de nos analyses. Dans la section qui suit, nous présentons quelques ajustements que nous avons apportés à cette méthode dans le but de réduire ses erreurs de prédiction.

## 5.6 Raffinement de la méthode exponentielle

Après avoir comparé les deux méthodes de prédiction, nous avons tenté d'améliorer la performance de la méthode sélectionnée en lui apportant de légères modifications. Nous ne voulions pas appliquer ces changements lors de la comparaison avec la méthode linéaire, car nous cherchions à déterminer quel type de fonction s'ajuste le mieux à l'ensemble des données historiques. Il fallait comparer les deux méthodes de façon équitable, donc il était préférable de ne pas appliquer des traitements qui risqueraient de favoriser l'une ou l'autre des méthodes. Une fois la méthode exponentielle choisie, notre objectif était de réduire au maximum les erreurs de prédiction qui y sont associées, ces erreurs étant quantifiées par la valeur des quatre critères présentés à la section précédente.

Repositionnement dynamique des fonctions de prédiction

Tel qu'expliqué à la section 5.4, les fonctions de prédiction sont positionnées (leur échelle de grandeur est déterminée) en calculant les valeurs $k_{clics}$ et $k_{cpc}$. Ces valeurs à l'origine sont calculées en fonction de $posMoyClics$, $posMoyCpc$, $clicsMoy$ et $cpcMoy$, qui sont obtenues à partir de l'historique de données. Initialement, nous calculions ces valeurs une seule fois afin d'obtenir une fonction de prédiction, puis nous utilisions cette fonction pour prédire toutes les journées subséquentes (30 jours de prédiction dans notre cas). Sachant que le milieu des publicités sur les moteurs de recherche est très dynamique et que plusieurs facteurs externes peuvent influencer la performance à court terme d'une campagne, nous croyons que le repositionnement dynamique des fonctions pourrait potentiellement diminuer les marges d'erreur associées aux prédictions.

Le Tableau 5.5 montre l'exemple d'un mot-clé pour lequel le repositionnement dynamique des fonctions de prédiction offrirait probablement des gains considérables. Ce mot-clé, tout comme plusieurs autres que nous avons rencontrés lors de nos analyses, a subi une baisse subite de son volume d'impressions et de clics quotidien pendant sa période de 30 jours de prédiction. Dans ce cas particulier, les valeurs de CPC moyen et de position moyenne sont demeurées relativement stables, donc la diminution de volume (observée à partir du 47$^{e}$ jour) peut être attribuée à des

facteurs externes. Les nombreux facteurs externes qui pourraient être responsables de cet effet sont expliqués à la section 4.1.

Tableau 5.5 : Données d'un mot-clé qui a subi une forte diminution de volume pendant sa période de prédiction

| Période historique (30 jours) | | | | | Période de prédiction (30 jours) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Jour | Position moyenne | Nombre d'impressions | Nombre de clics | CPC moyen | Jour | Position moyenne | Nombre d'impressions | Nombre de clics | CPC moyen |
| 1 | 2,89 | 251 | 12 | 0,204 | 31 | 2,86 | 1537 | 55 | 0,165 |
| 2 | 3,36 | 1480 | 37 | 0,188 | 32 | 2,91 | 1487 | 49 | 0,169 |
| 3 | 3,17 | 747 | 20 | 0,187 | 33 | 2,58 | 1913 | 84 | 0,195 |
| 4 | 2,99 | 1391 | 34 | 0,175 | 34 | 2,61 | 1892 | 66 | 0,190 |
| 5 | 3,03 | 1114 | 30 | 0,193 | 35 | 2,54 | 2109 | 98 | 0,197 |
| 6 | 2,96 | 745 | 23 | 0,186 | 36 | 2,58 | 2078 | 100 | 0,197 |
| 7 | 2,90 | 752 | 19 | 0,187 | 37 | 2,85 | 1644 | 46 | 0,188 |
| 8 | 2,89 | 735 | 29 | 0,204 | 38 | 2,88 | 1580 | 59 | 0,189 |
| 9 | 2,92 | 678 | 18 | 0,193 | 39 | 2,89 | 1392 | 52 | 0,184 |
| 10 | 3,01 | 688 | 23 | 0,200 | 40 | 2,89 | 1327 | 41 | 0,188 |
| 11 | 2,91 | 639 | 22 | 0,178 | 41 | 2,65 | 1164 | 56 | 0,185 |
| 12 | 2,31 | 813 | 34 | 0,193 | 42 | 2,76 | 1167 | 45 | 0,193 |
| 13 | 2,49 | 671 | 30 | 0,204 | 43 | 2,77 | 1343 | 54 | 0,183 |
| 14 | 2,85 | 641 | 18 | 0,138 | 44 | 2,94 | 1316 | 52 | 0,178 |
| 15 | 2,80 | 521 | 20 | 0,175 | 45 | 2,64 | 846 | 32 | 0,182 |
| 16 | 2,22 | 2239 | 103 | 0,189 | 46 | 2,80 | 739 | 30 | 0,183 |
| 17 | 2,77 | 2328 | 116 | 0,191 | 47 | 2,15 | 97 | 2 | 0,235 |
| 18 | 2,74 | 2187 | 107 | 0,186 | 48 | 2,76 | 54 | 3 | 0,200 |
| 19 | 2,95 | 801 | 23 | 0,178 | 49 | 2,14 | 51 | 2 | 0,230 |
| 20 | 2,97 | 761 | 23 | 0,193 | 50 | 1,84 | 56 | 2 | 0,180 |
| 21 | 2,52 | 1196 | 37 | 0,196 | 51 | 1,87 | 47 | 0 | N/A |
| 22 | 2,51 | 1103 | 37 | 0,199 | 52 | 2,61 | 44 | 1 | 0,170 |
| 23 | 2,27 | 1794 | 49 | 0,202 | 53 | 2,02 | 50 | 1 | 0,180 |
| 24 | 2,45 | 1070 | 48 | 0,188 | 54 | 2,17 | 30 | 1 | 0,200 |
| 25 | 2,18 | 1359 | 55 | 0,218 | 55 | 2,15 | 67 | 4 | 0,193 |
| 26 | 2,34 | 914 | 36 | 0,188 | 56 | 1,59 | 56 | 3 | 0,207 |
| 27 | 2,13 | 1668 | 64 | 0,216 | 57 | 2,17 | 104 | 5 | 0,192 |
| 28 | 2,26 | 1455 | 74 | 0,213 | 58 | 2,14 | 74 | 2 | 0,225 |
| 29 | 2,49 | 1778 | 78 | 0,196 | 59 | 2,21 | 67 | 3 | 0,217 |
| 30 | 2,62 | 1719 | 64 | 0,198 | 60 | 2,02 | 47 | 1 | 0,170 |

Dans le cas d'un tel mot-clé, l'algorithme de prédiction surestimera énormément les valeurs de clics prédites à partir de la 47$^{e}$ journée. Puisque les 30 jours d'historique utilisés avaient un volume moyen beaucoup plus élevé, la valeur de $k_{clics}$ sera beaucoup trop grande pour prédire les journées 47 à 60 (notons que la position n'a pas beaucoup varié, donc les fonctions de prédiction prédiront sensiblement les mêmes valeurs que ce qui a été observé pendant la période historique). En calculant cette valeur, nous obtenons :

$$k_{clics} = \frac{clicsMoy}{(c_{clics})^{posMoy}} = \frac{42{,}77}{(0{,}613)^{2{,}70}} = 160{,}08\ clics$$

Avec cette valeur initiale, les nombres de clics prédits à partir des positions moyennes des journées 47 à 60 sont inscrits au Tableau 5.6. Comme il est facile de constater, la baisse subite de volume du mot-clé fait en sorte que les valeurs de prédiction sont beaucoup trop élevées.

Tableau 5.6 : Prédictions de clics en fonction de la position moyenne pour les journées 47 à 60

| Jour | Position moyenne | Nombre de clics observés | Nombre de clics prédits |
|---|---|---|---|
| 47 | 2,15 | 2 | 55,77 |
| 48 | 2,76 | 3 | 41,48 |
| 49 | 2,14 | 2 | 56,24 |
| 50 | 1,84 | 2 | 65,07 |
| 51 | 1,87 | 0 | 64,03 |
| 52 | 2,61 | 1 | 44,55 |
| 53 | 2,02 | 1 | 59,57 |
| 54 | 2,17 | 1 | 55,44 |
| 55 | 2,15 | 4 | 55,92 |
| 56 | 1,59 | 3 | 73,54 |
| 57 | 2,17 | 5 | 55,27 |
| 58 | 2,14 | 2 | 56,30 |
| 59 | 2,21 | 3 | 54,31 |
| 60 | 2,02 | 1 | 59,53 |

Cet exemple démontre pourquoi il est important de considérer des données relativement récentes pour prédire le comportement des mots-clés. Nous avons donc choisi de tester l'effet du repositionnement dynamique des fonctions sur le potentiel de prédiction. Avec cette approche, de nouvelles valeurs de $k_{clics}$ et $k_{cpc}$ sont calculées à chaque journée de prédiction, en décalant la fenêtre de 30 journées historiques utilisées. Par exemple, la prédiction du 60$^{e}$ jour utilisera les 30 journées qui le précèdent, soit les jours 30 à 59.

Le Tableau 5.7 montre les résultats obtenus avec cette méthode en comparaison avec les résultats initiaux qui étaient fournis par la méthode exponentielle. Les résultats démontrent clairement que le repositionnement dynamique des fonctions de prédiction (l'utilisation d'une fenêtre mobile pour le calcul des valeurs initiales) apporte des gains considérables au niveau de l'exactitude des prédictions. Les valeurs d'erreur ont diminué pour chacun des quatre critères définis initialement. Nous en concluons donc que cette modification de la méthode devrait être conservée.

Tableau 5.7 : Comparaison des résultats de la méthode exponentielle sans repositionnement dynamique (A) et avec repositionnement dynamique (B)

| méthode | critère # 1 | critère # 2 | critère # 3 | critère # 4 |
|---|---|---|---|---|
| (A) | 0,5270 | 0,3066 | -0,0346 | -0,1058 |
| (B) | 0,4111 | 0,1965 | 0,0149 | -0,0184 |

Pondération des observations

Afin de complémenter le repositionnement dynamique des fonctions de prédiction, nous croyons qu'il pourrait être avantageux d'attribuer des poids différents à chacune des observations de la période historique (lors du calcul de $posMoyClics$, $posMoyCpc$, $clicsMoy$ et $cpcMoy$), dans le but de favoriser les observations les plus récentes. Comme l'a démontré l'exemple précédent, les mots-clés peuvent subir de fortes variations de volume en très peu de temps, ce qui a pour effet d'affecter la qualité des prédictions obtenues. Pour s'adapter à ces variations, nous proposons d'appliquer des poids exponentiellement décroissants aux observations de la période historique, en fonction de leur ancienneté. Voici la formule de calcul des poids que nous appliquons à nos données historiques :

$$poids_j = S^{30-j}$$

où $S$ est une constante comprise entre 0 et 1, $j$ est la journée considérée ($j$ = 1 à 30) et $poids_j$ est le poids que nous attribuons à l'observation de la journée $j$.

Pour appliquer ces poids aux observations, il suffit d'utiliser les formules suivantes lors du calcul des points moyens des graphiques :

$$posMoyClics = \frac{\sum_{j=1}^{n}(posClics_j * poids_j)}{\sum_{j=1}^{n} poids_j}$$

$$posMoyCpc = \frac{\sum_{j=1}^{m}(posCpc_j * poids_j)}{\sum_{j=1}^{m} poids_j}$$

$$clicsMoy = \frac{\sum_{j=1}^{n}(clics_j * poids_j)}{\sum_{j=1}^{n} poids_j}$$

$$cpcMoy = \frac{\sum_{j=1}^{m}(cpc_j * poids_j)}{\sum_{j=1}^{m} poids_j}$$

où $n$ est le nombre d'observations ayant au moins une impression dans la période historique (i.e. nombre de points sur le graphique de clics), $m$ est le nombre d'observations ayant au moins un clic dans la période historique (i.e. nombre de points sur le graphique de CPC moyen), $posClics_j$ est la position moyenne observée à la journée $j$ sur le graphique de clics, $posCpc_j$ est la position moyenne observée à la journée $j$ sur le graphique de CPC moyen, $clics_j$ est le nombre de clics observé à la journée $j$, $cpc_j$ est le CPC moyen observé à la journée $j$ et $poids_j$ est le poids que nous attribuons à l'observation de la journée $j$.

Nous avons testé plusieurs valeurs de $S$ afin d'étudier leur effet sur la qualité des prédictions obtenues. Évidemment, plus la valeur utilisée est petite, plus les poids attribués aux observations décroissent rapidement en fonction de leur ancienneté. Le Tableau 5.8 montre les différentes valeurs de poids qui sont calculées en fonction des valeurs de $S$ utilisées. Nous pouvons constater que certaines valeurs de $S$ font en sorte que l'historique de 30 jours n'est pas nécessaire, puisque les poids attribués aux observations deviennent quasi-nuls à partir d'un certain point.

Le Tableau 5.9 quantifie l'effet de l'ajout d'une pondération des observations sur les valeurs des quatre critères d'erreur, pour chacune des valeurs de $S$ considérées. Les résultats démontrent que plusieurs valeurs de $S$ semblent fournir de bons résultats. Il n'existe aucune solution dominante, comme le démontrent les taux d'erreur minimaux qui ont été mis en évidence dans chacune des colonnes du tableau. Afin de conserver une certaine stabilité dans nos prédictions, nous préférons éviter les valeurs de $S$ trop faibles. Il est important d'avoir une certaine sensibilité face aux changements de volumes des campagnes, mais nous ne voulons pas qu'une observation inhabituelle vienne complètement bouleverser les prédictions d'un mot-clé. Puisqu'il n'existe aucune solution clairement meilleure que les autres, nous choisissons arbitrairement la valeur $S = 0{,}70$, qui semble fournir des marges d'erreur acceptables selon chacun des quatre critères.

Comme le montre le Tableau 5.8, cette valeur nous permet de prioriser fortement les observations récentes, mais elle considère toutefois avec un poids non négligeable une bonne proportion des observations un peu plus anciennes.

Tableau 5.8 : Poids calculés en fonction des valeurs de *S* utilisées, pour une période de 30 jours

| jour | valeurs de S | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0,1 | 0,2 | 0,3 | 0,4 | 0,5 | 0,6 | 0,7 | 0,8 | 0,9 | 0,95 | 1 |
| 1 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,002 | 0,047 | 0,226 | 1,000 |
| 2 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,002 | 0,052 | 0,238 | 1,000 |
| 3 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,002 | 0,058 | 0,250 | 1,000 |
| 4 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,003 | 0,065 | 0,264 | 1,000 |
| 5 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,004 | 0,072 | 0,277 | 1,000 |
| 6 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,005 | 0,080 | 0,292 | 1,000 |
| 7 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,006 | 0,089 | 0,307 | 1,000 |
| 8 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,007 | 0,098 | 0,324 | 1,000 |
| 9 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,001 | 0,009 | 0,109 | 0,341 | 1,000 |
| 10 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,001 | 0,012 | 0,122 | 0,358 | 1,000 |
| 11 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,001 | 0,014 | 0,135 | 0,377 | 1,000 |
| 12 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,002 | 0,018 | 0,150 | 0,397 | 1,000 |
| 13 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,002 | 0,023 | 0,167 | 0,418 | 1,000 |
| 14 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,003 | 0,028 | 0,185 | 0,440 | 1,000 |
| 15 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,005 | 0,035 | 0,206 | 0,463 | 1,000 |
| 16 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,001 | 0,007 | 0,044 | 0,229 | 0,488 | 1,000 |
| 17 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,001 | 0,010 | 0,055 | 0,254 | 0,513 | 1,000 |
| 18 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,002 | 0,014 | 0,069 | 0,282 | 0,540 | 1,000 |
| 19 | 0,000 | 0,000 | 0,000 | 0,000 | 0,000 | 0,004 | 0,020 | 0,086 | 0,314 | 0,569 | 1,000 |
| 20 | 0,000 | 0,000 | 0,000 | 0,000 | 0,001 | 0,006 | 0,028 | 0,107 | 0,349 | 0,599 | 1,000 |
| 21 | 0,000 | 0,000 | 0,000 | 0,000 | 0,002 | 0,010 | 0,040 | 0,134 | 0,387 | 0,630 | 1,000 |
| 22 | 0,000 | 0,000 | 0,000 | 0,001 | 0,004 | 0,017 | 0,058 | 0,168 | 0,430 | 0,663 | 1,000 |
| 23 | 0,000 | 0,000 | 0,000 | 0,002 | 0,008 | 0,028 | 0,082 | 0,210 | 0,478 | 0,698 | 1,000 |
| 24 | 0,000 | 0,000 | 0,001 | 0,004 | 0,016 | 0,047 | 0,118 | 0,262 | 0,531 | 0,735 | 1,000 |
| 25 | 0,000 | 0,000 | 0,002 | 0,010 | 0,031 | 0,078 | 0,168 | 0,328 | 0,590 | 0,774 | 1,000 |
| 26 | 0,000 | 0,002 | 0,008 | 0,026 | 0,063 | 0,130 | 0,240 | 0,410 | 0,656 | 0,815 | 1,000 |
| 27 | 0,001 | 0,008 | 0,027 | 0,064 | 0,125 | 0,216 | 0,343 | 0,512 | 0,729 | 0,857 | 1,000 |
| 28 | 0,010 | 0,040 | 0,090 | 0,160 | 0,250 | 0,360 | 0,490 | 0,640 | 0,810 | 0,903 | 1,000 |
| 29 | 0,100 | 0,200 | 0,300 | 0,400 | 0,500 | 0,600 | 0,700 | 0,800 | 0,900 | 0,950 | 1,000 |
| 30 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 | 1,000 |

En retenant la solution $S = 0{,}70$ du Tableau 5.9, nous pouvons quantifier l'effet de l'ajout de la pondération des observations en comparant les valeurs de critère d'erreur avec celles obtenues précédemment. Les résultats du Tableau 5.10 démontrent clairement que l'application d'un poids aux observations en fonction de leur ancienneté diminue considérablement les erreurs associées à chacun des quatre critères. Il est donc évident que ce raffinement de notre méthode de prédiction doit être retenu.

Tableau 5.9 : Valeurs des critères d'erreur suite à l'application de divers poids exponentiels aux observations

| S | critère # 1 | critère # 2 | critère # 3 | critère # 4 |
|---|---|---|---|---|
| 0,10 | 0,3386 | 0,1775 | **0,0008** | -0,0169 |
| 0,20 | 0,3354 | 0,1743 | 0,0012 | -0,0168 |
| 0,30 | 0,3336 | 0,1717 | 0,0017 | -0,0167 |
| 0,40 | **0,3333** | 0,1696 | 0,0021 | -0,0166 |
| 0,50 | 0,3342 | 0,1680 | 0,0026 | -0,0164 |
| 0,60 | 0,3360 | 0,1669 | 0,0033 | -0,0163 |
| 0,70 | 0,3395 | **0,1666** | 0,0043 | **-0,0162** |
| 0,80 | 0,3471 | 0,1684 | 0,0062 | **-0,0162** |
| 0,90 | 0,3669 | 0,1760 | 0,0102 | -0,0167 |
| 0,95 | 0,3861 | 0,1844 | 0,0129 | -0,0174 |
| 1,00 | 0,4111 | 0,1965 | 0,0149 | -0,0184 |

Tableau 5.10 : Comparaison des résultats de la méthode exponentielle sans repositionnement dynamique (A), avec repositionnement dynamique (B) et avec repositionnement dynamique ainsi que pondération des observations (C)

| méthode | critère # 1 | critère # 2 | critère # 3 | critère # 4 |
|---|---|---|---|---|
| **(A)** | 0,5270 | 0,3066 | -0,0346 | -0,1058 |
| **(B)** | 0,4111 | 0,1965 | 0,0149 | -0,0184 |
| **(C)** | 0,3395 | 0,1666 | 0,0043 | -0,0162 |

## 5.7 Conclusion

Récapitulation de la méthode

Au cours de notre recherche de la meilleure méthode de prédiction générique, nous avons analysé et comparé plusieurs approches. D'abord, nous avons considéré des fonctions linéaires ainsi que des fonctions exponentielles. Afin de comparer le potentiel de chacune de ces méthodes pour prédire les clics et les CPC moyens en fonction de la position moyenne, nous avons dû développer des critères objectifs. En comparant les résultats obtenus, nous avons conclu que les deux méthodes avaient des performances très semblables, mais que la méthode exponentielle fournissait toujours des résultats légèrement meilleurs. De plus, celle-ci présentait plus d'avantages que la méthode linéaire au niveau des facteurs non quantifiables. Bref, nous avons opté pour une approche de prédiction générique qui utilise des fonctions de prédiction exponentielles, autant dans le cas des clics que pour les CPC moyens.

Par la suite, une fois le meilleur type de fonction identifié, nous avons essayé d'appliquer plusieurs raffinements à notre méthode de prédiction dans le but de réduire les taux d'erreur obtenus. Suite à plusieurs tests, nous avons conclu que le repositionnement dynamique des fonctions ainsi que la pondération des observations historiques selon leur ancienneté permettait de réduire considérablement la valeur de chacun des quatre critères d'erreur identifiés. Nous avons donc choisi d'inclure ces raffinements dans notre méthode de prédiction des clics et des CPC moyens.

Commentaire sur la qualité des résultats obtenus

Grâce à tous les raffinements qui ont été appliqués à la méthode de prédiction, les taux d'erreur ont atteint des valeurs de 0,3395 et 0,1666 pour les critères 1 et 2 respectivement. De plus, les erreurs associées aux critères 3 et 4 ont atteint des valeurs presque nulles, ce qui implique que globalement, les fonctions génériques ne surestiment et ne sous-estiment pas les valeurs totales de clics et de coûts sur la période de 30 jours étudiée.

Évidemment, nous sommes satisfaits des résultats fournis par les critères 3 et 4. Cependant, à première vue, il peut être difficile de déterminer si les valeurs atteintes par les critères 1 et 2 sont satisfaisantes. En effet, il existe toujours une certaine dispersion dans les données qui fait en sorte

que la valeur minimale pouvant être atteinte par les critères 1 et 2 n'est pas nulle. Ainsi, la valeur de référence à utiliser pour évaluer la qualité des taux d'erreur obtenus pour ces critères doit être prise en compte lors de l'évaluation de la qualité des résultats.

Afin de présenter ce concept à l'aide d'un exemple, considérons le mot-clé illustré à la Figure 5.10. Nous jugeons que les régressions de clics et de CPC moyens de ce mot-clé sont adéquates (et même très satisfaisantes) et qu'elles peuvent donc être utilisées pour modéliser son comportement en fonction de la position moyenne. En effet, il est facile de constater que les observations suivent une tendance décroissante bien définie et qu'elles sont bien regroupées autour des courbes de régression. Les valeurs des critères d'erreur associées à ces fonctions de régression, inscrites au Tableau 5.11, nous permettent de comparer ce cas avec les résultats fournis par les fonctions génériques. L'exemple démontre que même dans les meilleurs cas (i.e. les mots-clés à faible dispersion dans les données, qui peuvent être modélisés avec des régressions satisfaisantes), les valeurs d'erreur calculées pour les critères 1 et 2 sont non négligeables. Même si la méthode des régressions est censée fournir le meilleur ajustement sur les observations (selon le critère des moindres carrés), les taux d'erreur obtenus sont de 0,1453 pour le critère 1 et 0,1040 pour le critère 2.

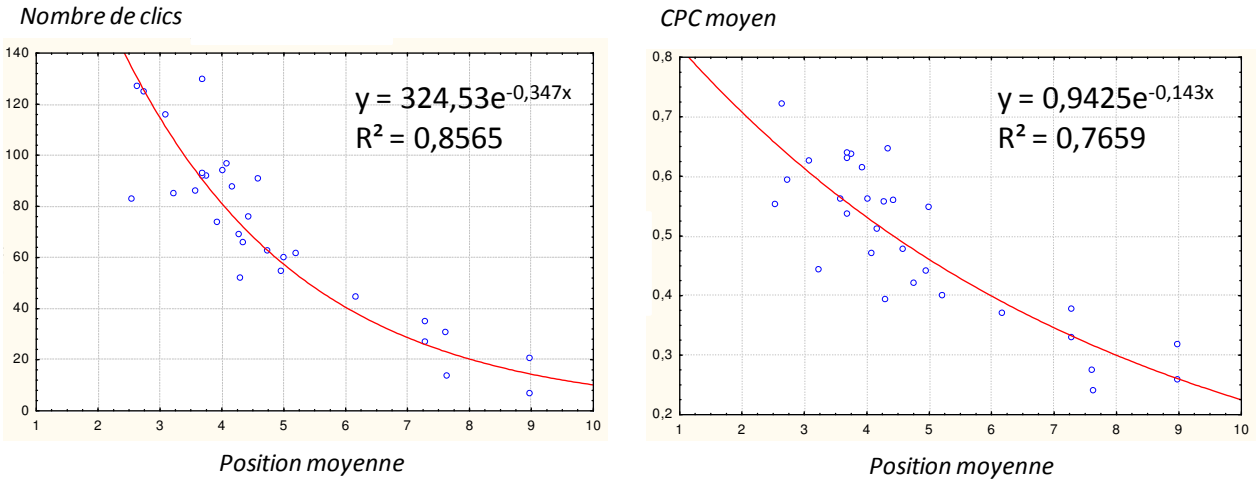



Figure 5.10 : Exemple de mot-clé pour lequel la régression fournit une fonction de prédiction satisfaisante

Tableau 5.11 : Valeurs des critères d'erreur obtenues pour le mot-clé de la Figure 5.10 en utilisant les régressions comme fonctions de prédiction

| Critère | Valeur |
|---|---|
| 1 | 0,1453 |
| 2 | 0,1040 |
| 3 | -0,0103 |
| 4 | -0,0126 |

En comparant les valeurs d'erreur calculées pour l'ensemble de nos fonctions génériques avec celles obtenues pour un mot-clé que nous considérons comme idéal, il devient plus facile de comprendre ce que représentent ces taux d'erreur. Compte tenu de ces résultats, nous sommes généralement satisfaits de la qualité des prédictions fournies par les fonctions génériques. À notre avis, les marges d'erreur obtenues lors des premières phases de tests étaient beaucoup trop élevées, mais elles diminuaient à mesure que nous ajoutions des raffinements à notre méthode. Considérant tous les facteurs difficiles à modéliser ou même imprévisibles qui peuvent influencer la performance des campagnes publicitaires, nous jugeons que les seuils d'erreur atteints jusqu'à présent sont acceptables.

Les méthodes de prédiction génériques pourraient donc, selon nous, être utilisées dans le cadre du modèle d'optimisation, dans le but d'obtenir des estimations de clics et de CPC moyen pour les mots-clés qui ne fournissent pas des régressions satisfaisantes. Nous avons plusieurs idées qui permettraient probablement de réduire davantage les marges d'erreur associées aux prédictions, mais, faute de temps, nous n'avons pas pu les appliquer. Afin d'encourager la poursuite de la recherche dans cette direction, nous présenterons certaines pistes de recherche intéressantes au chapitre suivant.

# CONCLUSION

Récapitulation

Le monde des annonces textuelles sur les moteurs de recherche est tellement vaste et complexe qu'il est presque nécessaire d'utiliser des logiciels et outils d'aide à la décision pour obtenir des retours sur investissement satisfaisants. Les campagnes publicitaires étant typiquement constituées de plusieurs dizaines de milliers de mots-clés, il est impossible de les gérer adéquatement sans avoir recours à diverses stratégies de gestion automatisées. Considérant que les données historiques associées à chacun de ces mots-clés sont accessibles quotidiennement, il est facile de reconnaître les gains potentiels associés à l'utilisation de méthodes de statistique et recherche opérationnelle. C'est pourquoi ce milieu, relativement nouveau, commence à susciter l'intérêt de plusieurs professionnels et chercheurs du milieu académique.

L'objectif principal de notre projet était de modéliser et prédire le comportement des mots-clés dans des campagnes d'annonces textuelles sur les moteurs de recherche, dans le but de développer des algorithmes automatisables qui permettront d'améliorer le rendement de ces campagnes. Plus précisément, nous cherchions à exploiter l'information fournie par les données historiques des mots-clés afin de permettre aux utilisateurs de la plateforme logicielle d'Acquisio de gérer leurs valeurs d'enchère de la façon la plus intelligente possible.

Pour y arriver, il fallut d'abord comprendre le fonctionnement des mécanismes publicitaires en détail. Puisque le domaine est relativement très peu connu et mal documenté, cette étape fut particulièrement difficile. Nous avons dû investir beaucoup de temps afin d'acquérir une bonne compréhension des divers éléments qui constituent l'environnement publicitaire à l'étude, ainsi que les effets de leurs interactions.

Par la suite, nous avons consulté plusieurs références relatives à nos objectifs de recherche. Étant donné la nouveauté du domaine et la forte compétition entre les entreprises qui développent des logiciels de gestion d'enchères, il n'existe que très peu de documentation pertinente publiée à ce jour. Plusieurs des références considérées exploraient des approches intéressantes et ont inspiré certaines idées de modélisation présentées dans ce mémoire, mais aucune d'entre elles ne semblait offrir une méthode répondant à tous nos besoins. Ces recherches ont donc permis de

valider qu'il n'existait pas, dans la littérature, une version équivalente aux algorithmes que nous cherchions à développer. De plus, elles nous ont fourni des pistes qui ont permis de mieux orienter notre travail.

Nous avons alors poursuivi notre démarche en définissant un programme linéaire d'optimisation nous permettant d'atteindre des objectifs publicitaires très spécifiques, tout en tenant compte des contraintes relatives au contexte publicitaire. À partir de cette modélisation, nous avons déterminé qu'il était nécessaire de procéder à une classification des mots-clés afin de leur affecter des traitements différents, en fonction de leurs caractéristiques spécifiques. Puis, nous avons également dû développer des méthodes de prédiction alternatives afin d'augmenter considérablement la proportion des mots-clés pouvant être pris en compte par le modèle d'optimisation. Étant donné leur importance critique, nous avons consacré une grande partie de nos efforts au développement et à l'amélioration des fonctions de prédiction génériques.

Les contributions du mémoire

Premièrement, nous avons réussi à décortiquer et organiser l'information pertinente relative au domaine de façon à la présenter clairement sous forme d'un résumé, qui pourrait être utilisé pour initier des nouveaux chercheurs au monde des annonces textuelles sur les moteurs de recherche. Il existe très peu de documentation dans la littérature qui présente de façon aussi exhaustive et structurée toute l'information nécessaire à la compréhension du domaine.

En second lieu, nous considérons que la modélisation proposée permet de mieux comprendre les interactions et dépendances entre chacune des variables du problème. En représentant le problème sous cet angle, nous avons réussi à le simplifier tout en ne perdant pas de vue les objectifs des annonceurs. De plus, nos choix de modélisation nous ont permis de mieux cibler nos besoins en termes de méthodes de classification et de prédiction nécessaires pour les accompagner.

Les classifications que nous avons effectuées à partir de nos banques de données ont permis de comprendre la structure des campagnes publicitaires et plus spécifiquement les caractéristiques des mots-clés qui les constituent. Initialement, nous pensions qu'il serait possible d'appliquer des méthodes de régression à presque tous les mots-clés d'une campagne publicitaire. En explorant plus en profondeur, nous avons constaté que cette approche théoriquement idéale ne serait

applicable que dans une faible proportion des cas. Nous avons alors conclu qu'il serait nécessaire de développer des options de prédiction alternatives afin de maximiser l'étendue du modèle d'optimisation. Bref, nos essais de classification ont démontré à quel point les campagnes publicitaires sont complexes et nous ont permis d'ajuster nos méthodes de prédiction en fonction des diverses tendances observées.

Dans cette optique, nous avons orienté nos efforts de recherche envers l'élaboration de méthodes de prédiction génériques. L'idée derrière cette approche peut sembler relativement simple, mais il était difficile de l'appliquer en pratique. Il fallut constamment ajuster les algorithmes pour tenir compte de nombreux cas d'exception, afin que la méthode puisse être testée sur des banques de données de plusieurs centaines de milliers de mots-clés. Nous avons subi plusieurs échecs avant d'arriver aux méthodes présentées dans ce mémoire. Les nombreux tests effectués ont toutefois permis d'approfondir nos connaissances relatives au comportement des mots-clés et à certaines particularités du problème. Après avoir apporté diverses améliorations à notre méthode, nous avons finalement réussi à atteindre des résultats que nous jugeons très satisfaisants. À notre avis, les marges d'erreur obtenues démontraient que les prédictions étaient relativement précises, considérant la généralité de la méthode. Cependant, plusieurs pistes de recherche demeurent inexplorées et nous croyons qu'il serait possible d'améliorer davantage la qualité des prédictions obtenues en approfondissant certains concepts.

Pistes de recherche à explorer

Nous avons opté pour une modélisation de type sac-à-dos binaire à choix multiple avec des prédictions de clics et de CPC moyen en fonction de la position moyenne, mais il serait évidemment possible d'évaluer le potentiel de plusieurs autres modèles semblables. Entre autres, puisque les fonctions de prédiction que nous utilisons sont des fonctions continues, il serait possible de définir un modèle à variables de position continues. Dans un tel cas, les positions moyennes ciblées pourraient prendre des valeurs non entières et les contraintes d'intégrité seraient donc éliminées du modèle. Cependant, il faudrait évaluer l'effet d'une telle modification sur le temps de résolution du programme, car elle rendrait le modèle non linéaire.

Il serait également possible de modifier le modèle de façon à utiliser des fonctions qui prédisent le nombre de clics en fonction du CPC moyen (ou de la valeur d'enchère, si celle-ci est

disponible dans l'historique de données). Une telle approche éliminerait la nécessité d'aborder le concept de position moyenne, représentant plutôt le nombre de clics obtenu en fonction du montant déboursé par clic. De cette façon, il faudrait une variable intermédiaire de moins pour prédire les coûts associés à chacune des positions potentielles et la performance globale de l'algorithme risquerait d'en bénéficier. Dans le cadre de ce mémoire, nous avons orienté nos recherches autour de fonctions de prédiction qui utilisent la position comme variable indépendante, car cela semblait être l'approche la plus communément utilisée dans la littérature. Il serait toutefois intéressant de voir si les régressions et les fonctions génériques fournissent de bons résultats dans le cas du nombre de clics en fonction du CPC moyen (ou la valeur d'enchère).

Par ailleurs, il serait intéressant de tenir compte de la périodicité des volumes de requêtes associés aux mots-clés pour ajuster les fonctions de prédiction. Nous savons que certaines campagnes publicitaires et plus spécifiquement certains mots-clés subissent de fortes variations de volume, que ce soit à une fréquence annuelle, mensuelle, hebdomadaire ou quotidienne. Par exemple, un magasin qui vend des accessoires de jardinage risque d'obtenir beaucoup plus d'impressions et de clics en période estivale. D'un autre côté, un site Web qui offre des promotions dans des restaurants risque d'être plus recherché la fin de semaine. Les observations recueillies dans la base de données ne sont donc pas toutes équivalentes; il n'est pas souhaitable de comparer, sur un graphique, une observation obtenue en période achalandée avec une autre observation recueillie lors d'une période moins occupée. Il faudrait donc évaluer si l'ajout d'un poids à chacune des observations permettrait d'obtenir de meilleures prédictions, autant lors du calcul des régressions que dans le cas de fonctions génériques (par exemple, pour un mot-clé qui subit des variations de volume sur une base hebdomadaire, attribuer des poids différents à chacune des journées de la semaine). Dans certains cas, il serait peut-être même possible d'utiliser des séries temporelles pour identifier les variations cycliques.

Il pourrait aussi être avantageux de considérer l'impact de plusieurs variables indépendantes lors du calcul des régressions. Actuellement, nous utilisons des régressions simples qui prédisent la valeur des clics ou des CPC moyens en fonction d'une seule variable : la position moyenne. Cependant, il existe probablement d'autres facteurs qui ont un effet sur la valeur de la variable indépendante. Entre autres, nous croyons que des régressions multiples pourraient être utilisées en tenant compte de variables telles que le type de correspondance associé au mot-clé, le moteur

de recherche sur lequel le mot-clé se retrouve, le nombre total de termes distincts compris dans le mot-clé, le type de produit associé au mot-clé, l'indice de qualité du mot-clé, etc.

Dans le cas où une des variables mentionnées aurait une influence significative sur les prédictions, il pourrait également être intéressant de vérifier l'effet de celle-ci sur les taux de décroissance génériques des mots-clés. En effet, nous croyons que certaines caractéristiques pourraient contribuer à déterminer la décroissance relative d'un mot-clé. En effectuant des analyses sur des ensembles de mots-clés partageant des caractéristiques communes, il serait possible de déterminer lesquelles des caractéristiques exercent une influence importante sur les valeurs génériques des coefficients de décroissance. Plus spécifiquement, nous proposons de reproduire une analyse semblable à celle présentée aux Figures 5.7 et 5.8, en mettant en évidence les valeurs des diverses variables à l'étude. En évaluant l'effet des variables de cette façon, il serait peut-être possible d'attribuer des valeurs génériques différentes à certains ensembles de mots-clés, en se basant sur leurs caractéristiques particulières. Par exemple, un test qui pourrait être effectué serait de comparer les coefficients de décroissance des mots-clés en fonction de leur type de correspondance (voir section 1.2 pour explication des types de correspondance). Si le test démontre que les types de correspondance ont des taux de décroissance différents pour leurs fonctions de clics et de CPC moyen, il serait possible d'exploiter cette information pour mieux prédire le comportement des mots-clés dans le futur.

Finalement, il faudrait développer davantage les habiletés d'adaptation des fonctions de prédiction génériques. En effet, les fonctions génériques ne fournissent que des estimations et nous sommes conscients que celles-ci risquent de s'éloigner considérablement des valeurs observées dans certains cas. Afin d'exploiter au maximum l'information à notre disposition, il faudrait trouver des méthodes permettant d'ajuster constamment les prédictions en fonction des nouvelles observations recueillies. Par exemple, il serait avantageux d'ajuster les coefficients de décroissance utilisés pour prédire le comportement d'un mot-clé si les données récentes démontrent qu'il ne se comporte pas tel que prévu.

En conclusion, le problème d'optimisation des campagnes d'annonces textuelles est loin d'être résolu. Notre recherche a mené à certaines découvertes intéressantes, mais nous n'avons pas encore assemblé chacun des aspects du projet de façon à constituer un algorithme d'optimisation complet et fonctionnel. Il reste à déterminer si l'utilisation de nos méthodes de prédiction en

parallèle avec le modèle d'optimisation proposé permettront réellement d'améliorer le rendement des campagnes publicitaires des annonceurs. Il faudra approfondir certains concepts et procéder à plusieurs séries de tests avant de pouvoir appliquer nos méthodes de prédiction et d'optimisation à des campagnes publicitaires réelles.

# RÉFÉRENCES

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

Agarwal, A., Hosanagar, K., & Smith, M. D. (2008). *Location, Location, Location: An Analysis of Profitability of Position in Online Advertising Markets* (56). Pittsburgh, États-Unis: Heinz Research. Consulté le 19 janvier 2011, tiré de http://repository.cmu.edu/cgi/viewcontent.cgi?article=1055&context=heinzworks&sei-redir=1#search=%22Location,+Location,+Location,+An+Analysis+of+Profitability+of+Position+in+Online+Advertising+Markets%22

Agichtein, E., Brill, E., Dumais, S., & Ragno, R. (2006). Learning User Interaction Models for Predicting Web Search Result Preferences [Version électronique]. *SIGIR 2006: Conference on Research and Development in Information Retrieval*, (pp. 3-10). New York, États-Unis: ACM.

Borgs, C., Chayes, J., Etesami, O., Immorlica, N., Jain, K., & Mahdian, M. (2007). Dynamics of Bid Optimization in Online Advertisement Auctions [Version électronique]. *WWW2007: International World Wide Web Conference*, (pp. 531-540). New York, États-Unis: ACM.

Brooks, N. (2004a). *The Atlas Rank Report : How Search Engine Rank Impacts Trafic*. The Atlas Institute. Consulté le 19 mai 2009, tiré de http://www.atlassolutions.com/uploadedFiles/Atlas/Atlas_Institute/Published_Content/RankReport.pdf

Brooks, N. (2004b). *The Atlas Rank Report – Part II : How Search Engine Rank Impacts Conversions*. The Atlas Institute. Consulté le 19 mai 2009, tiré de http://surf2your.pages.com.au/resources/RankReportPart2.pdf

Chakrabarty, D., Zhou, Y., & Lukose, R. (2008). Budget Constrained Bidding in Keyword Auctions and Online Knapsack Problems [Version électronique]. *WINE2008: Workshop on Internet and Network Economics*, (pp. 566-576). Shanghai, Chine: Springer-Verlag Berlin Heidelberg.

Craswell, N., Zoeter, O., Taylor, M., & Ramsey, B. (2008). An Experimental Comparison of Click Position-Bias Models [Version électronique]. *WSDM 2008 : International Conference on Web Search and Data Mining*, (pp. 87-94). New York, États-Unis: ACM.

Edelman, B., Ostrovsky, M., & Schwarz, M. (2006). Internet Advertising and the Generalized Second-Price Auction: Selling Billions of Dollars Worth of Keywords. *American Economic Review, 97*(1), 242-259. Consulté le 19 mai 2009, tiré de http://rwj.berkeley.edu/schwarz/publications/gsp051003.pdf

Even Dar, E., Mirrokni, V. S., Muthukrishnan, S., Mansour, Y., & Nadav, U. (2009). Bid Optimization for Broad Match Ad Auctions [Version électronique]. *WWW2009: International World Wide Web Conference*, (pp. 231-240). New York, États-Unis: ACM.

Feldman, J., & Muthukrishnan, S. (2008). Algorithmic Methods for Sponsored Search Advertising. In Liu, Z., & Xia, C. H., Performance Modeling and Engineering (pp. 91-122). Hawthorne, NY, États-Unis: Springer US. Consulté le 28 mars 2011, tiré de springerlink.com.

Feldman, J., Muthukrishnan, S., Pal, M., & Stein, C. (2008). Budget Optimization in Search-Based Advertising Auctions [Version électronique]. *EC 2007: ACM Conference on Electronic Commerce,* (pp. 40-49). New York, États-Unis: ACM.

Ganchev, K., Kulesza, A., Tan, J., Gabbard, R., Liu, Q., & Kearns, M. (2007). Empirical Price Modeling for Sponsored Search [Version électronique]. *WINE2007: Workshop on Internet and Network Economics*, (pp. 541-548). San Diego, États-Unit: Springer-Verlag Berlin Heidelberg.

Google Finance (2010a). *Google Inc. Financials*. Google Finance. Consulté le 7 mars 2011, tiré de http://www.google.com/finance?q=NASDAQ:GOOG&fstype=ii

Google Finance (2010b). *Yahoo! Inc. Financials*. Google Finance. Consulté le 7 mars 2011, tiré de http://www.google.com/finance?q=NASDAQ:YHOO&fstype=ii

Google Finance (2010c). *Baidu.com, Inc. Financials*. Google Finance. Consulté le 7 mars 2011, tiré de http://www.google.com/finance?q=NASDAQ:BIDU&fstype=ii

Google Inc. (2008). *2008 Annual Report*. Google Inc. Consulté le 7 mars 2011, tiré de investor.google.com/pdf/2008_google_annual_report.pdf

Hines, W., Montgomery, D. C., Goldsman, D. M., Borror, C. M., Adjengue, L.-D., & Carmichael, J.-P. (2005). Probabilités et statistique pour ingénieurs (1ère éd.). Montréal, Canada: Les Éditions de la Chenelière inc.

Hotchkiss, G., Alston, S., & Edwards, G. (2005). Google Eye Tracking Report. Enquiro Research. Consulté le 19 mai 2009, tiré de http://pages.enquiro.com/whitepaper-enquiro-eye-tracking-report-I-google.html

Hou, L., Wang, L., & Yang, J. (2008). Evolutionary Prediction of Online Keywords Bidding [Version électronique]. *EC-Web 2008: 9th International Conference of Electronic Commerce and Web Technologies,* (pp. 124-133). Turin, Italie: Springer-Verlag Berlin Heidelberg.

International Telecommunication Union (2010). *Market Information and Statistics*. International Telecommunication Union. Consulté le 7 mars 2011, tiré de http://www.itu.int/ITU-D/ict/statistics/

Jansen, B. J., & Mullen, T. (2008). Sponsored Search : An Overview of the Concept, History and Technology. *International Journal of Electronic Business, 6*(2), 114-131. Consulté le 19 mai 2009, tiré de http://faculty.ist.psu.edu/jjansen/academic/pubs/jansen_overview_sponosored_search.pdf

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately Interpreting Clickthrough Data as Implicit Feedback [Version électronique]. *SIGIR 2005: Conference on Research and Development in Information Retrieval*, (pp. 154-161). New York, États-Unis: ACM.

Kempe, D., & Mahdian, M. (2008). A Cascade Model for Externalities in Sponsored Search [Version électronique]. *WINE2008: Workshop on Internet and Network Economics*, (pp. 585-596). Shanghai, Chine: Springer-Verlag Berlin Heidelberg.

Kitts, B., & Leblanc, B. (2004). Optimal Bidding on Keyword Auctions [Version électronique]. *Electronic Markets, 14*(3), 186-201.

Kitts, B., Laxminarayan, P., Leblanc, B., & Meech, R. (2005). A Formal Analysis of Search Auctions Including Predictions on Click Fraud and Bidding Tactics [Version électronique]. *EC 2005: ACM Conference on Electronic Commerce*. New York, États-Unis: ACM.

Liu, D., Chen, J., & Whinston, A. B. (2009). Current Issues in Keyword Auctions. In Adomavicius, G., & Gupta, A., Handbooks in Information Systems (Vol. 3, pp. 69-97). Bingley, Royaume-Uni: Emerald Group Publishing Limited. Consulté le 28 mars 2011, tiré de Business Computing.

Martello, S., & Toth, P. (1990). Knapsack Problems: Algorithms and Computer Implementations. Chichester, Angleterre: John Wiley & Sons Ltd. Consulté le 4 avril 2011, tiré de Università di Bologna - D.E.I.S. - Operations Research: http://www.or.deis.unibo.it/kp/KnapsackProblems.pdf

Microsoft Corporation (2009). *Microsoft Corporation 2009 Annual Report : Financial Review*. Microsoft Corporation. Consulté le 7 mars 2011, tiré de http://www.microsoft.com/investor/reports/ar09/10k_fr_not_21.html

Microsoft Encarta (2011). *Encyclopédie Encarta en ligne*. Consulté le 7 mars 2011, tiré de http://fr.encarta.msn.com/encnet/features/dictionary/DictionaryResults.aspx?lextype=3&search=moteur%20de%20recherche

Muthukrishnan, S., Pal, M., & Svitkina, Z. (2010). Stochastic Models for Budget Optimization in Search-Based Advertising. *Algorithmica, 58*(4), 1022-1044. Consulté le 28 mars 2011, tiré de http://www.springerlink.com/content/42k0021140937438/

netmarketshare.com (2011). *Usage Share Statistics for Internet Technologies: Search Engine Market Share*. Consulté le 28 mars 2011, tiré de http://www.netmarketshare.com/search-engine-market-share.aspx?qprid=4

Rusmevichientong, P., & Williamson, D. P. (2006). An Adaptive Algorithm for Selecting Profitable Keywords for Search-Based Advertising Services [Version électronique]. *EC 2006: ACM Conference on Electronic Commerce,* (pp. 260-269). New York, États-Unis: ACM.

Schonfeld, E. (2010). *Forecast: Online Retail Sales Will Grow to $250 Billion by 2014*. Consulté le 7 mars 2011, tiré de http://seekingalpha.com/article/192498-forecast-online-retail-sales-will-grow-to-250-billion-by-2014

Varian, H. (2009). *Conversion Rates Don't Vary Much with Ad Position*. Google Inside AdWords. Consulté le 7 mars 2011, tiré de http://adwords.blogspot.com/2009/08/conversion-rates-dont-vary-much-with-ad.html

# ANNEXE 1 – Répartition des taux de clic moyens en fonction de la position moyenne pour les 20 banques de données testées

| | BANQUE DE DONNÉES #1 | | | BANQUE DE DONNÉES #2 | | |
|---|---|---|---|---|---|---|
| position | Nombre de clics | Nombre d'impressions | Taux de clic moyen | Nombre de clics | Nombre d'impressions | Taux de clic moyen |
| 1 | 10479487 | 241187145 | 0,0434 | 4957431 | 68053515 | 0,0728 |
| 2 | 4958141 | 230058775 | 0,0216 | 3187574 | 141441920 | 0,0225 |
| 3 | 2397573 | 141768158 | 0,0169 | 3041748 | 206825942 | 0,0147 |
| 4 | 1201253 | 94778624 | 0,0127 | 2564121 | 225332126 | 0,0114 |
| 5 | 596772 | 58389061 | 0,0102 | 1701781 | 188315444 | 0,0090 |
| 6 | 317043 | 36064135 | 0,0088 | 1011405 | 134301320 | 0,0075 |
| 7 | 139998 | 19283593 | 0,0073 | 512041 | 76138245 | 0,0067 |
| 8 | 64071 | 11670971 | 0,0055 | 254566 | 39511717 | 0,0064 |
| 9 | 23829 | 5729755 | 0,0042 | 113955 | 19369772 | 0,0059 |
| 10 | 10758 | 2909542 | 0,0037 | 52222 | 10027428 | 0,0052 |

| | BANQUE DE DONNÉES #3 | | | BANQUE DE DONNÉES #4 | | |
|---|---|---|---|---|---|---|
| position | Nombre de clics | Nombre d'impressions | Taux de clic moyen | Nombre de clics | Nombre d'impressions | Taux de clic moyen |
| 1 | 4743198 | 264326910 | 0,0179 | 1359140 | 115763282 | 0,0117 |
| 2 | 2339819 | 417816142 | 0,0056 | 802081 | 48707042 | 0,0165 |
| 3 | 656133 | 94807632 | 0,0069 | 172180 | 17445924 | 0,0099 |
| 4 | 205033 | 32399356 | 0,0063 | 35288 | 4479667 | 0,0079 |
| 5 | 72236 | 15715260 | 0,0046 | 10796 | 1677031 | 0,0064 |
| 6 | 26929 | 7873849 | 0,0034 | 4239 | 759779 | 0,0056 |
| 7 | 10379 | 3167824 | 0,0033 | 1259 | 316154 | 0,0040 |
| 8 | 4230 | 1531580 | 0,0028 | 642 | 209912 | 0,0031 |
| 9 | 1624 | 639964 | 0,0025 | 165 | 126259 | 0,0013 |
| 10 | 740 | 323123 | 0,0023 | 70 | 53539 | 0,0013 |

| | BANQUE DE DONNÉES #5 | | | BANQUE DE DONNÉES #6 | | |
|---|---|---|---|---|---|---|
| position | Nombre de clics | Nombre d'impressions | Taux de clic moyen | Nombre de clics | Nombre d'impressions | Taux de clic moyen |
| 1 | 3494873 | 154233154 | 0,0227 | 143743 | 8046865 | 0,0179 |
| 2 | 4135522 | 168291229 | 0,0246 | 179811 | 18032773 | 0,0100 |
| 3 | 1629242 | 106185947 | 0,0153 | 51715 | 10084116 | 0,0051 |
| 4 | 556439 | 48766581 | 0,0114 | 14252 | 4800316 | 0,0030 |
| 5 | 167381 | 20990571 | 0,0080 | 4741 | 2262115 | 0,0021 |
| 6 | 51726 | 10146846 | 0,0051 | 1278 | 865078 | 0,0015 |
| 7 | 21287 | 5556282 | 0,0038 | 437 | 457513 | 0,0010 |
| 8 | 10051 | 2967501 | 0,0034 | 193 | 264945 | 0,0007 |
| 9 | 3118 | 995508 | 0,0031 | 62 | 119188 | 0,0005 |
| 10 | 917 | 369840 | 0,0025 | 29 | 41576 | 0,0007 |

| | BANQUE DE DONNÉES #7 | | | BANQUE DE DONNÉES #8 | | |
|---|---|---|---|---|---|---|
| position | Nombre de clics | Nombre d'impressions | Taux de clic moyen | Nombre de clics | Nombre d'impressions | Taux de clic moyen |
| 1 | 973421 | 118103467 | 0,0082 | 291894 | 8303337 | 0,0352 |
| 2 | 1283208 | 68305119 | 0,0188 | 149986 | 8653188 | 0,0173 |
| 3 | 414050 | 44201624 | 0,0094 | 48617 | 4125797 | 0,0118 |
| 4 | 93555 | 24597921 | 0,0038 | 13880 | 1738228 | 0,0080 |
| 5 | 22969 | 6480588 | 0,0035 | 3667 | 833249 | 0,0044 |
| 6 | 7849 | 1766936 | 0,0044 | 1103 | 321779 | 0,0034 |
| 7 | 3429 | 879737 | 0,0039 | 411 | 142714 | 0,0029 |
| 8 | 2158 | 2117249 | 0,0010 | 138 | 58748 | 0,0023 |
| 9 | 1029 | 1246161 | 0,0008 | 67 | 24423 | 0,0027 |
| 10 | 500 | 282402 | 0,0018 | 14 | 9416 | 0,0015 |

| | BANQUE DE DONNÉES #9 | | | BANQUE DE DONNÉES #10 | | |
|---|---|---|---|---|---|---|
| position | Nombre de clics | Nombre d'impressions | Taux de clic moyen | Nombre de clics | Nombre d'impressions | Taux de clic moyen |
| 1 | 1379920 | 144560892 | 0,0095 | 249855 | 7004774 | 0,0357 |
| 2 | 657031 | 167067147 | 0,0039 | 161999 | 13068007 | 0,0124 |
| 3 | 275277 | 40667736 | 0,0068 | 26009 | 3501491 | 0,0074 |
| 4 | 98011 | 17037715 | 0,0058 | 5708 | 1139301 | 0,0050 |
| 5 | 28815 | 5806743 | 0,0050 | 1228 | 297295 | 0,0041 |
| 6 | 10589 | 2595547 | 0,0041 | 93 | 21713 | 0,0043 |
| 7 | 4139 | 1171450 | 0,0035 | 12 | 2643 | 0,0045 |
| 8 | 1441 | 602949 | 0,0024 | 1 | 1117 | 0,0009 |
| 9 | 773 | 289556 | 0,0027 | 2 | 471 | 0,0042 |
| 10 | 281 | 223497 | 0,0013 | 2 | 278 | 0,0072 |

| | BANQUE DE DONNÉES #11 | | | BANQUE DE DONNÉES #12 | | |
|---|---|---|---|---|---|---|
| position | Nombre de clics | Nombre d'impressions | Taux de clic moyen | Nombre de clics | Nombre d'impressions | Taux de clic moyen |
| 1 | 374664 | 15770944 | 0,0238 | 39571 | 1137879 | 0,0348 |
| 2 | 220598 | 25001429 | 0,0088 | 4965 | 393057 | 0,0126 |
| 3 | 52607 | 10811953 | 0,0049 | 1197 | 236412 | 0,0051 |
| 4 | 16067 | 3792329 | 0,0042 | 829 | 170540 | 0,0049 |
| 5 | 4599 | 980725 | 0,0047 | 407 | 102170 | 0,0040 |
| 6 | 2200 | 529550 | 0,0042 | 36 | 31556 | 0,0011 |
| 7 | 987 | 301593 | 0,0033 | 10 | 12144 | 0,0008 |
| 8 | 515 | 196245 | 0,0026 | 6 | 6458 | 0,0009 |
| 9 | 191 | 102723 | 0,0019 | 1 | 2501 | 0,0004 |
| 10 | 116 | 71648 | 0,0016 | 0 | 297 | 0,0000 |

| | BANQUE DE DONNÉES #13 | | | BANQUE DE DONNÉES #14 | | |
|---|---|---|---|---|---|---|
| position | Nombre de clics | Nombre d'impressions | Taux de clic moyen | Nombre de clics | Nombre d'impressions | Taux de clic moyen |
| 1 | 270980 | 5196635 | 0,0521 | 88014 | 68841840 | 0,0013 |
| 2 | 43408 | 2984063 | 0,0145 | 25662 | 9804259 | 0,0026 |
| 3 | 4618 | 486115 | 0,0095 | 7990 | 2144567 | 0,0037 |
| 4 | 655 | 100025 | 0,0065 | 982 | 217572 | 0,0045 |
| 5 | 154 | 31726 | 0,0049 | 138 | 38387 | 0,0036 |
| 6 | 108 | 16236 | 0,0067 | 42 | 10115 | 0,0042 |
| 7 | 49 | 7281 | 0,0067 | 8 | 4114 | 0,0019 |
| 8 | 33 | 7024 | 0,0047 | 8 | 1905 | 0,0042 |
| 9 | 26 | 3960 | 0,0066 | 5 | 850 | 0,0059 |
| 10 | 7 | 1390 | 0,0050 | 1 | 494 | 0,0020 |

| | BANQUE DE DONNÉES #15 | | | BANQUE DE DONNÉES #16 | | |
|---|---|---|---|---|---|---|
| position | Nombre de clics | Nombre d'impressions | Taux de clic moyen | Nombre de clics | Nombre d'impressions | Taux de clic moyen |
| 1 | 27260 | 46672374 | 0,0006 | 124216 | 3679049 | 0,0338 |
| 2 | 8146 | 7734865 | 0,0011 | 194925 | 12489227 | 0,0156 |
| 3 | 2146 | 2166285 | 0,0010 | 72716 | 9726408 | 0,0075 |
| 4 | 474 | 471896 | 0,0010 | 18695 | 5736303 | 0,0033 |
| 5 | 183 | 338921 | 0,0005 | 5762 | 2461155 | 0,0023 |
| 6 | 189 | 275109 | 0,0007 | 2276 | 1254312 | 0,0018 |
| 7 | 49 | 125040 | 0,0004 | 958 | 655396 | 0,0015 |
| 8 | 15 | 13205 | 0,0011 | 556 | 455633 | 0,0012 |
| 9 | 12 | 52724 | 0,0002 | 258 | 232420 | 0,0011 |
| 10 | 1 | 941 | 0,0011 | 120 | 114076 | 0,0011 |

| | BANQUE DE DONNÉES #17 | | | BANQUE DE DONNÉES #18 | | |
|---|---|---|---|---|---|---|
| position | Nombre de clics | Nombre d'impressions | Taux de clic moyen | Nombre de clics | Nombre d'impressions | Taux de clic moyen |
| 1 | 37165259 | 454775643 | 0,0817 | 4616858 | 573804731 | 0,0080 |
| 2 | 29148604 | 596506133 | 0,0489 | 347188 | 30966758 | 0,0112 |
| 3 | 18592191 | 576124824 | 0,0323 | 91520 | 11821093 | 0,0077 |
| 4 | 9076847 | 354599789 | 0,0256 | 36886 | 7036507 | 0,0052 |
| 5 | 2624703 | 140087122 | 0,0187 | 17344 | 3893859 | 0,0045 |
| 6 | 810240 | 58923338 | 0,0138 | 9175 | 2461611 | 0,0037 |
| 7 | 297800 | 28135307 | 0,0106 | 4517 | 1500508 | 0,0030 |
| 8 | 107714 | 13709606 | 0,0079 | 2670 | 1052345 | 0,0025 |
| 9 | 33989 | 5825758 | 0,0058 | 1460 | 678203 | 0,0022 |
| 10 | 13078 | 2700980 | 0,0048 | 812 | 466380 | 0,0017 |

| | BANQUE DE DONNÉES #19 | | | BANQUE DE DONNÉES #20 | | |
|---|---|---|---|---|---|---|
| position | Nombre de clics | Nombre d'impressions | Taux de clic moyen | Nombre de clics | Nombre d'impressions | Taux de clic moyen |
| 1 | 43311448 | 205235933 | 0,2110 | 1702658 | 170381330 | 0,0100 |
| 2 | 12080315 | 331292696 | 0,0365 | 999922 | 61315316 | 0,0163 |
| 3 | 7924972 | 380126194 | 0,0208 | 495250 | 74552406 | 0,0066 |
| 4 | 4624519 | 303076516 | 0,0153 | 166623 | 33517690 | 0,0050 |
| 5 | 2304997 | 206001236 | 0,0112 | 67785 | 16706201 | 0,0041 |
| 6 | 1118983 | 118585384 | 0,0094 | 26149 | 7912155 | 0,0033 |
| 7 | 555362 | 67176149 | 0,0083 | 10665 | 4171438 | 0,0026 |
| 8 | 323597 | 42031974 | 0,0077 | 3346 | 1542290 | 0,0022 |
| 9 | 172468 | 24340141 | 0,0071 | 1247 | 1058355 | 0,0012 |
| 10 | 78592 | 12348380 | 0,0064 | 629 | 947112 | 0,0007 |

# ANNEXE 2 – Mécanisme de diversification des positions

Tel qu'expliqué à la section 4.1, plusieurs stratégies de gestion de campagnes publicitaires ont pour effet de faire stabiliser les valeurs d'enchère à des niveaux fixes pour des périodes de temps prolongées. En conséquence, les mots-clés demeurent à des positions relativement constantes d'une journée à l'autre. Avec des étendues de position très restreintes, il est difficile d'obtenir des régressions qui peuvent capturer l'effet de décroissance des clics et des CPC moyens en fonction de la position. En effet, lorsque les campagnes sont gérées d'une telle façon, nous possédons très peu d'informations quant au potentiel d'amélioration du rendement de la campagne dans les autres positions et il n'est pas possible d'obtenir une affectation optimale des mots-clés aux positions en utilisant le modèle d'optimisation.

Afin d'acquérir le plus d'information possible concernant la variation des valeurs de clics et CPC moyen, il peut être avantageux de faire varier les positions visitées par un mot-clé. C'est dans ce but que nous avons décidé de développer un mécanisme de diversification des positions. En modifiant stratégiquement les valeurs d'enchères d'un mot-clé, il serait possible de faire varier les positions occupées d'une journée à l'autre de façon à maximiser le nombre de positions différentes qui sont visitées. Nous croyons que l'élargissement de l'étendue de positions aura un impact positif sur la qualité des régressions obtenues.

La Figure A-2.1 illustre un exemple de l'effet de la diversification des positions sur un mot-clé. Avec les graphiques initiaux (A et B), il est très difficile de prédire le nombre de clics ou les CPC moyens qui seront obtenus à des positions autres que 2, 3 et 4. En effet, les 30 observations se regroupent sous forme de nuages de points sans tendance et l'application de régressions (linéaires dans cet exemple) sur les données fournit des prédictions très peu fiables. En observant le graphique B, nous constatons que les CPC moyens sont demeurés relativement stables d'une journée à l'autre, ce qui nous pousse à croire que les valeurs d'enchère n'ont pas beaucoup varié. Cela pourrait donc expliquer pourquoi l'étendue de positions est aussi restreinte. Lorsque 30 jours de données additionnels avec positions variées sont ajoutés aux graphiques (C et D), l'effet de la position sur les clics et les CPC moyens devient plus évident. Il est alors possible d'effectuer des prédictions à l'aide de régressions.

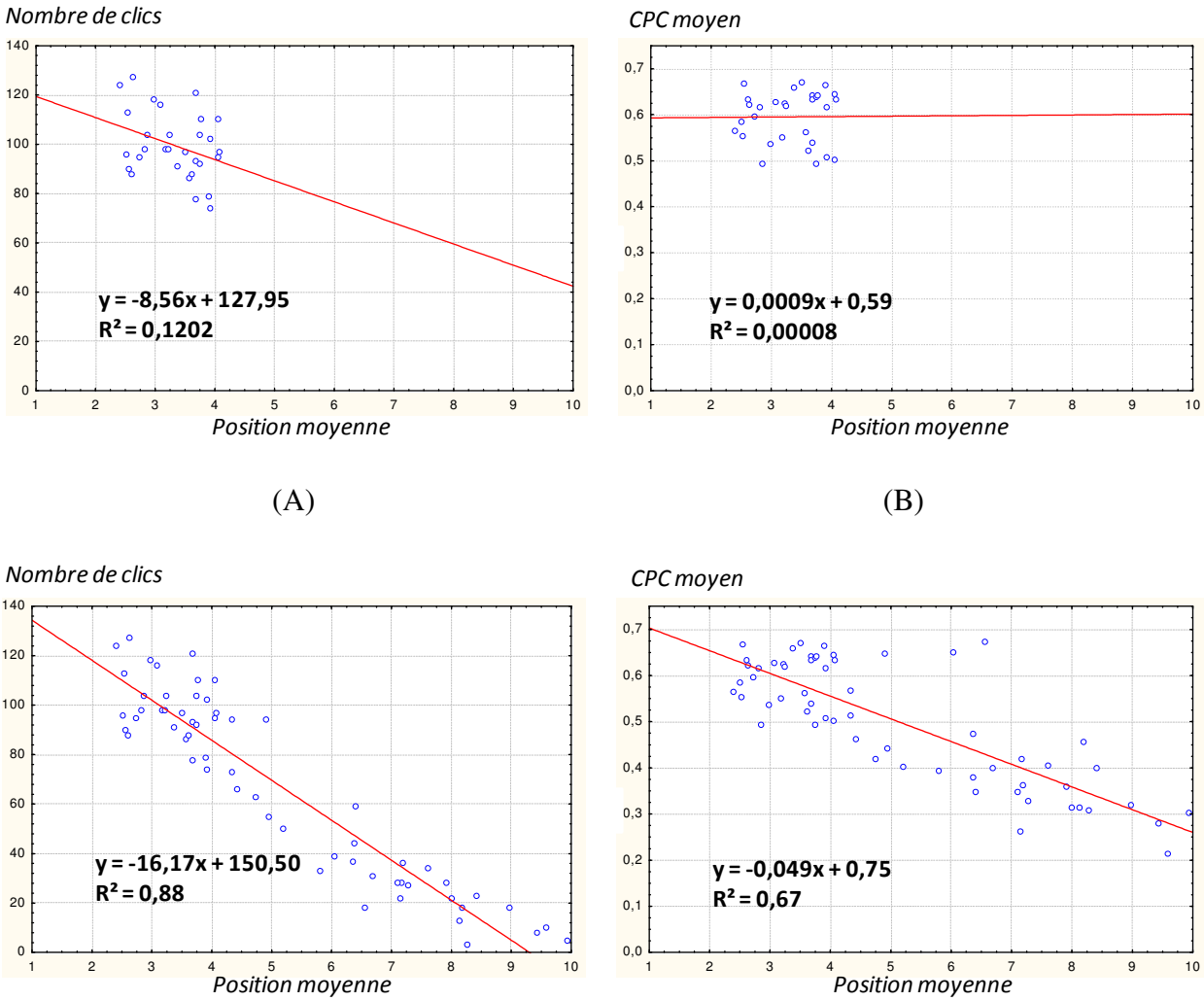



Figure A-2.1 : Exemples de fonctions de clics et de CPC moyens avant (A, B) et après (C, D) diversification des positions

Premiers tests

Afin de valider l'hypothèse que la diversification des positions permet d'améliorer la qualité des fonctions de prédiction, nous avons décidé d'effectuer quelques tests sur un ensemble de mots-clés réels. Avec l'accord d'un client, nous avons sélectionné un groupe de quelques mots-clés pour lesquels nous avions la liberté de faire varier les valeurs d'enchère à chaque jour.

Après plus d'un mois d'expérimentation, nous avons finalement conclu que la variation des enchères avait un impact direct sur la qualité des prédictions obtenues. Lorsque les valeurs d'enchère demeuraient sensiblement les mêmes, aucune tendance n'était apparente dans les données. Cependant, lorsque nous fixions les enchères de façon à augmenter considérablement l'étendue de positions visitées, l'effet de la position sur les clics, les CPC moyens et les valeurs d'enchère devenait de plus en plus apparent. Plus les enchères étaient uniformément réparties entre chacune des positions non visitées, plus les effets de tendance devenaient faciles à détecter.

Pour poursuivre notre analyse, nous avons également décidé d'effectuer quelques tests à l'aide de notre base de données afin d'analyser l'effet de la diversification des positions sur la qualité des fonctions de régression obtenues. En définissant des critères de recherche qui exigeaient une étendue de positions minimale et une bonne répartition de ces positions, nous avons réussi à extraire plusieurs ensembles de mots-clés différents de notre base de données. Plus précisément, nous avons décidé de quantifier l'ampleur de la diversification des positions selon deux critères.

1) Étendue : Mesurée par la différence entre la position maximale et la position minimale du mot-clé.

2) Répartition : Mesurée par l'écart-type de toutes les valeurs de position du mot-clé.

De plus, afin d'extraire uniquement des mots-clés qui ont des nombres suffisants de clics et d'observations pour potentiellement générer des régressions de qualité, nous avons choisi d'imposer deux autres critères.

3) Volume de clics : Mesuré par le nombre moyen de clics par jour du mot-clé.

4) Nombre d'observations : Mesuré par le nombre total d'observations disponibles pour le mot-clé.

Tout comme à la section 4.2, nous avons choisi d'évaluer la qualité des régressions obtenues selon les critères du coefficient de détermination $R^2$ et des conditions nécessaires. Nous avons constaté que la proportion des mots-clés qui satisfaisaient nos critères de qualité augmentait

constamment lorsque les deux critères de diversification devenaient plus sévères (i.e. lorsque les valeurs des seuils minimaux augmentaient), autant pour les fonctions de clics que pour les fonctions de CPC moyens.

Le Tableau A-2.1 montre l'effet de la variation des seuils pour les critères d'étendue et de répartition des positions (critères 1 et 2) sur la qualité des régressions obtenues, pour un exemple effectué à partir d'une banque de données possédant plus de 310 000 mots-clés. Pour chacun des scénarios testés, le nombre de mots-clés satisfaisant nos critères de qualité ($R^2$ minimal atteint et conditions nécessaires satisfaites avec régression exponentielle ou linéaire) est comptabilisé.

Tableau A-2.1 : Proportion de mots-clés satisfaisant les conditions de qualité en fonction des valeurs des critères de répartition de la position

| Seuils minimaux | | | | Résultats | |
|---|---|---|---|---|---|
| Étendue ($pos_{max}$ - $pos_{min}$) | Répartition (écart-type) | Volume de clics | Nombre d'observations | Nombre de mots-clés obtenus | Proportion satisfaisant les critères de qualité |
| 2,0 | 0,5 | 10 | 100 | 812 | **68,1%** |
| 3,0 | 1,0 | 10 | 100 | 516 | **81,8%** |
| 4,0 | 1,5 | 10 | 100 | 267 | **87,3%** |
| 5,0 | 2,0 | 10 | 100 | 85 | **95,3%** |

Cet exemple met en évidence plusieurs faits intéressants concernant la diversification des positions. D'abord, il montre à quel point les mots-clés à positions variées sont rares; par exemple, il existe seulement 85 mots-clés qui ont une étendue de plus de 5 positions avec un écart-type de plus de 2 sur leurs valeurs de position, tout en satisfaisant les conditions minimales de volume et de nombre d'observations. Cela ne représente que 0,027% du nombre total de mots-clés contenus dans la banque de données. De plus, l'exemple permet de constater l'effet de l'étendue et la répartition des positions observées sur la qualité des fonctions de prédiction; à mesure que les critères d'étendue et de répartition des positions deviennent plus sévères, la proportion du nombre de mots-clés satisfaisant nos critères de qualité augmente. La qualité des prédictions est donc beaucoup influencée par l'étendue de positions d'un mot-clé.

Devant ces résultats, nous sommes arrivés à la conclusion que le choix stratégique des valeurs d'enchère, en considérant les positions déjà occupées, aurait un effet positif sur la qualité des prédictions obtenues. Le problème consisterait donc à trouver une façon d'automatiser ce choix

« stratégique » des valeurs d'enchère, afin de l'appliquer quotidiennement à des ensembles de plusieurs centaines de mots-clés. L'objectif d'un tel algorithme serait essentiellement de fournir un maximum d'observations pertinentes tout en limitant les coûts et le temps nécessaires pour obtenir cette information.

Présentation de l'algorithme

À première vue, l'algorithme décrit précédemment peut sembler relativement facile à créer. Effectivement, il serait possible d'utiliser une simple fonction qui génère des positions au hasard et qui leur associe des valeurs d'enchère. Cependant, une telle méthode ne fournirait pas un rendement optimal. Lorsqu'on vise à maximiser le nombre de positions différentes obtenues tout en respectant des contraintes de coût et de temps, la complexité du problème devient apparente. Afin d'explorer les différentes positions de la façon la plus efficace possible, il faut tenir compte des observations passées et de leur effet sur les prédictions. Suite à plusieurs essais, nous sommes finalement arrivés à un algorithme de diversification des positions qui prend en considération toutes les spécificités du problème.

Pour des raisons de confidentialité, les détails mathématiques du modèle ne sont pas mentionnés. Cependant, le fonctionnement global de l'algorithme est présenté et le rôle de chacun des éléments au sein de celui-ci est expliqué.

Ensemble de mots-clés considéré par l'algorithme

Les mots-clés qui ont été affectés au mécanisme de diversification des positions feront partie de l'ensemble à traiter. Cet ensemble peut représenter quelques mots-clés ou plusieurs milliers de mots-clés, dépendant des critères de tri qui sont utilisés entre les étapes *G* et *H* de l'algorithme de classification. Généralement, le nombre de mots-clés affectés au mécanisme de diversification des positions ne devrait représenter qu'une faible proportion des mots-clés d'une campagne publicitaire.

Fréquence de résolution

L'algorithme est utilisé quotidiennement. Puisque les données fournies par les moteurs de recherche sont agrégées sur une base quotidienne, cette fréquence de résolution permet de maximiser le nombre d'observations différentes obtenues. À chaque jour, le mécanisme de diversification des positions détermine la valeur d'enchère qui doit être attribuée à chacun des mots-clés de l'ensemble à traiter.

Objectif

Pour chaque mot-clé, il faut déterminer le poids que nous associons au fait de visiter chaque position potentielle. En fonction des données historiques déjà disponibles, certaines positions auront des poids plus élevés que d'autres. Puisque nous cherchons à maximiser le nombre de positions différentes visitées par un mot-clé, ce sont les positions les moins fréquemment obtenues jusqu'à présent qui seront le plus valorisées. Par exemple, dans les graphiques A et B de la Figure A-2.1, les positions 1, 5, 6, 7, 8, 9 et 10 auraient un plus grand poids que les positions 2, 3 et 4, car elles n'ont pas encore été visitées. En effet, afin d'acquérir un maximum d'information concernant la décroissance des clics et des CPC moyens en fonction de la position, il serait important de visiter le plus grand nombre de positions possible.

Des poids différents sont calculés pour chacune des positions en fonction du nombre d'observations qu'elles possèdent, du nombre d'observations que leurs positions voisines possèdent et de l'ancienneté de ces observations. Il est très important de favoriser les observations les plus récentes, car le milieu étudié est très dynamique et les volumes peuvent varier énormément d'une période à l'autre. De plus, en accordant plus d'importance aux observations récentes, il est possible d'exploiter au maximum l'information fournie par l'algorithme de diversification des positions lors des journées précédentes. Une fois le calcul du poids de chaque position effectué, notre objectif consistera à maximiser la somme de ces poids lors de l'affectation des mots-clés à des positions.

Contrainte de budget

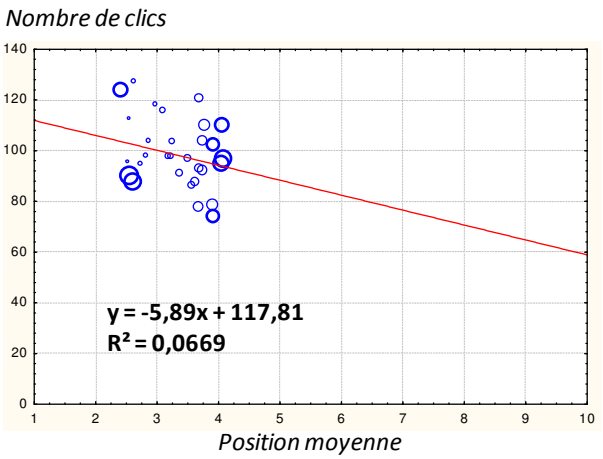
Lors de l'affectation des mots-clés aux différentes positions, il faut tenir compte d'une contrainte de budget pour l'ensemble de mots-clés à traiter. Afin de ne pas affecter les coûts totaux de la campagne publicitaire, nous choisissons d'accorder comme budget le même montant que ce qui était consommé par ces mots-clés avant leur entrée dans le mécanisme de diversification. Il suffit donc d'accorder comme budget la valeur moyenne des frais publicitaires associés aux mots-clés en question, sur la période historique considérée.

Prédictions

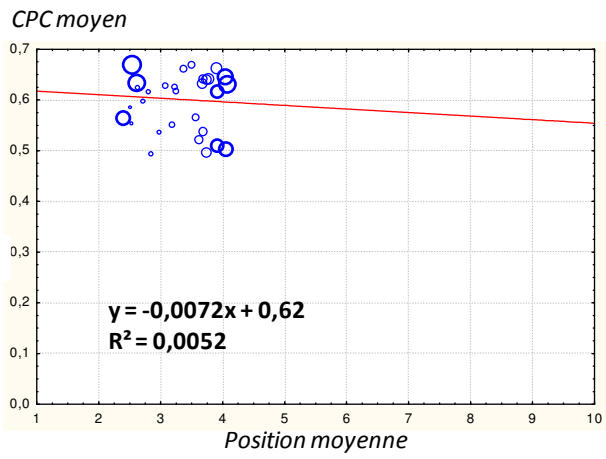
Puisque les mots-clés soumis au mécanisme de diversification des positions ne sont initialement pas capables de fournir des régressions qui satisfont nos critères de qualité, nous sommes conscients qu'il ne sera pas possible d'extrapoler les valeurs de clics et CPC moyens à partir des données disponibles. Ainsi, pour les premiers jours, il faudra d'abord générer des valeurs d'enchère aléatoires dans un intervalle de valeurs précis, afin d'obtenir des observations dans des positions variées et améliorer la qualité des prédictions de clics et CPC moyens en fonction de la position. Pendant ces journées, il sera difficile d'estimer les coûts totaux et le budget quotidien risque de ne pas toujours être respecté. Cependant, à mesure que de nouvelles observations seront recueillies, les fonctions de prédiction devraient se préciser. En considérant que nous attribuons une plus grande importance aux observations récentes, quelques jours devraient être suffisants pour préciser les fonctions de prédiction. Par la suite, il sera possible d'utiliser celles-ci pour estimer les coûts quotidiens des mots-clés et les valeurs d'enchère associées à chaque position.

Expliquons ce principe plus en détail à l'aide d'un exemple, illustré à la Figure A-2.2. Supposons que nous désirons effectuer des prédictions pour le mot-clé de la Figure A-2.1 (A et B). Puisque les régressions obtenues pour ce mot-clé ne permettent pas de prédire adéquatement son comportement, les valeurs d'enchère sont choisies de façon aléatoire pendant les trois premiers jours. Essentiellement, nous choisissons des valeurs au hasard dans un intervalle de valeurs que nous considérons raisonnable, tout en essayant d'éviter de répéter des valeurs d'enchère qui ont déjà été utilisées dans le passé. En fixant ces valeurs d'enchère, nous obtenons des positions moyennes de 8,41, 8,19 et 6,37 pour les trois premiers jours. Lorsque nous ajoutons ces observations au graphique, nous sommes en mesure d'estimer, par interpolation ou extrapolation,

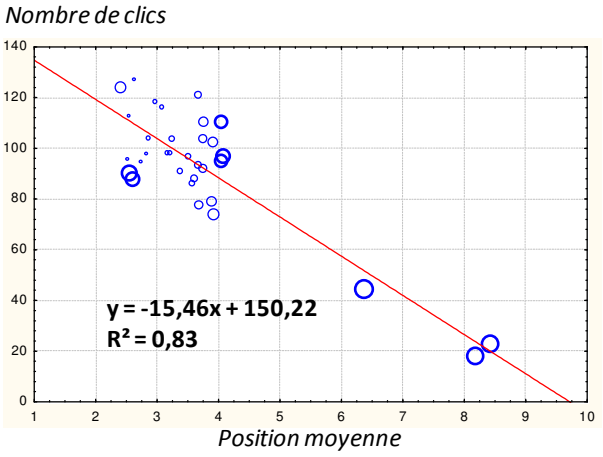
le comportement des clics et des CPC moyens pour une plus grande plage de positions. Notons qu'une pondération exponentiellement décroissante a été attribuée aux observations en fonction de leur ancienneté (représentée par la grosseur des points sur les graphiques), ce qui fait en sorte que les observations les plus récentes ont beaucoup d'effet sur la forme de la fonction de régression.



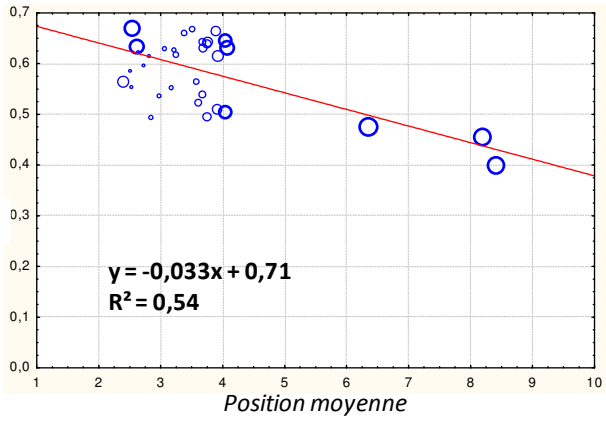

(A)




(B)




(C)




(D)

Figure A-2.2 : Effet de l'ajout d'observations avec valeurs d'enchères aléatoires sur les graphiques de clics pondérés (A et C) et de CPC moyens pondérés (B et D) en fonction de la position moyenne

Détails de l'algorithme

Une fois les premières valeurs obtenues, il est possible de poursuivre avec l'algorithme de diversification des positions. Tel que mentionné, nous tenterons de prioriser les positions les moins fréquemment visitées afin de maximiser l'information obtenue. À l'aide d'un programme linéaire d'optimisation, nous viserons à affecter une seule position à chacun des mots-clés de façon à maximiser la valeur associée aux positions choisies, tout en respectant une contrainte de budget. Les dépenses seront estimées, pour chacune des positions potentielles, à l'aide des fonctions de nombre de clics et CPC moyen (exemple : Figure A-2.2, C et D).

Au début, les fonctions obtenues par régression seront plus ou moins fiables et risquent de générer des erreurs de prédiction considérables. Cependant, avec l'acquisition de nouvelles observations dans des positions variées, leur qualité devrait augmenter graduellement, pour finalement converger vers une fonction de prédiction fiable. Dans certains cas, des prédictions de qualité ne pourront jamais être obtenues. Il faut donc fixer une durée maximale accordée à ce processus, afin d'éviter son utilisation pendant des périodes de temps trop prolongées. La Figure A-2.3 illustre le processus itératif au sein duquel l'algorithme de diversification des positions agit. À chaque jour, l'algorithme doit associer une valeur d'enchère à chaque mot-clé sélectionné pour la diversification, en se basant sur les données historiques actuellement disponibles. Le processus se répète quotidiennement, jusqu'à ce que les critères de qualité des fonctions soient satisfaits ou que la durée maximale initialement fixée soit atteinte.

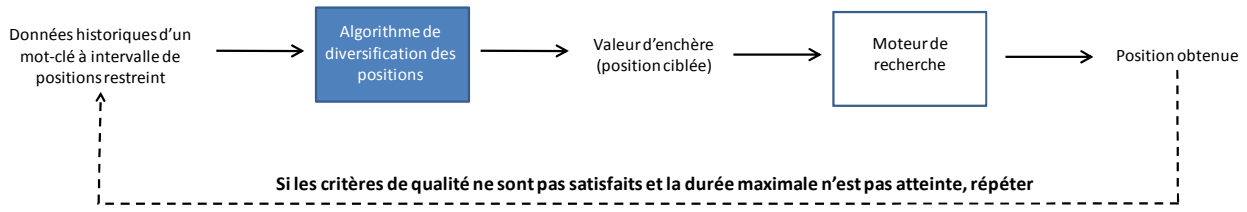



Figure A-2.3 : Rôle de l'algorithme de diversification des positions au sein du processus global

# ANNEXE 3 – Répartition des clics et des coûts dans les banques de données étudiées

| BANQUE DE DONNÉES #1 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 91 277,79 | 79,34% | 91 362,33 | 85,61% |
| 90-95 | 8 244,84 | 7,17% | 6 798,78 | 6,37% |
| 85-90 | 4 624,05 | 4,02% | 3 243,02 | 3,04% |
| 80-85 | 3 192,49 | 2,77% | 1 890,87 | 1,77% |
| 75-80 | 2 369,66 | 2,06% | 1 209,46 | 1,13% |
| 70-75 | 1 781,67 | 1,55% | 820,18 | 0,77% |
| 65-70 | 1 337,59 | 1,16% | 570,75 | 0,53% |
| 60-65 | 977,62 | 0,85% | 390,86 | 0,37% |
| 55-60 | 674,46 | 0,59% | 250,83 | 0,24% |
| 50-55 | 412,66 | 0,36% | 140,07 | 0,13% |
| 45-50 | 154,29 | 0,13% | 42,87 | 0,04% |
| 40-45 | 0,00 | 0,00% | 0,00 | 0,00% |
| 35-40 | 0,00 | 0,00% | 0,00 | 0,00% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #2 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 78 667,12 | 80,32% | 81 397,22 | 84,11% |
| 90-95 | 7 542,28 | 7,70% | 6 692,19 | 6,92% |
| 85-90 | 4 167,52 | 4,26% | 3 351,40 | 3,46% |
| 80-85 | 2 728,48 | 2,79% | 2 074,44 | 2,14% |
| 75-80 | 1 901,27 | 1,94% | 1 373,44 | 1,42% |
| 70-75 | 1 329,85 | 1,36% | 912,86 | 0,94% |
| 65-70 | 896,39 | 0,92% | 578,86 | 0,60% |
| 60-65 | 542,36 | 0,55% | 315,59 | 0,33% |
| 55-60 | 161,45 | 0,16% | 76,10 | 0,08% |
| 50-55 | 0,00 | 0,00% | 0,00 | 0,00% |
| 45-50 | 0,00 | 0,00% | 0,00 | 0,00% |
| 40-45 | 0,00 | 0,00% | 0,00 | 0,00% |
| 35-40 | 0,00 | 0,00% | 0,00 | 0,00% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #3 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 58 334,07 | 84,85% | 47 969,09 | 85,26% |
| 90-95 | 4 270,42 | 6,21% | 3 574,63 | 6,35% |
| 85-90 | 2 104,04 | 3,06% | 1 692,03 | 3,01% |
| 80-85 | 1 317,33 | 1,92% | 1 019,71 | 1,81% |
| 75-80 | 902,95 | 1,31% | 688,03 | 1,22% |
| 70-75 | 647,21 | 0,94% | 480,15 | 0,85% |
| 65-70 | 463,14 | 0,67% | 340,69 | 0,61% |
| 60-65 | 327,54 | 0,48% | 239,42 | 0,43% |
| 55-60 | 222,32 | 0,32% | 157,47 | 0,28% |
| 50-55 | 133,02 | 0,19% | 87,66 | 0,16% |
| 45-50 | 31,11 | 0,05% | 16,52 | 0,03% |
| 40-45 | 0,00 | 0,00% | 0,00 | 0,00% |
| 35-40 | 0,00 | 0,00% | 0,00 | 0,00% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #4 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 20 626,75 | 82,03% | 16 413,43 | 86,00% |
| 90-95 | 1 902,61 | 7,57% | 1 214,87 | 6,37% |
| 85-90 | 827,17 | 3,29% | 511,95 | 2,68% |
| 80-85 | 515,87 | 2,05% | 296,63 | 1,55% |
| 75-80 | 371,10 | 1,48% | 196,57 | 1,03% |
| 70-75 | 268,13 | 1,07% | 141,47 | 0,74% |
| 65-70 | 201,16 | 0,80% | 105,03 | 0,55% |
| 60-65 | 152,20 | 0,61% | 75,91 | 0,40% |
| 55-60 | 112,28 | 0,45% | 55,16 | 0,29% |
| 50-55 | 80,34 | 0,32% | 37,63 | 0,20% |
| 45-50 | 53,30 | 0,21% | 23,40 | 0,12% |
| 40-45 | 28,11 | 0,11% | 11,54 | 0,06% |
| 35-40 | 6,75 | 0,03% | 2,38 | 0,01% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #5 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 38 462,65 | 84,65% | 24 427,98 | 87,63% |
| 90-95 | 3 040,21 | 6,69% | 1 643,65 | 5,90% |
| 85-90 | 1 334,24 | 2,94% | 666,35 | 2,39% |
| 80-85 | 778,76 | 1,71% | 361,84 | 1,30% |
| 75-80 | 528,81 | 1,16% | 230,33 | 0,83% |
| 70-75 | 385,90 | 0,85% | 164,15 | 0,59% |
| 65-70 | 290,35 | 0,64% | 121,51 | 0,44% |
| 60-65 | 217,20 | 0,48% | 91,73 | 0,33% |
| 55-60 | 161,52 | 0,36% | 68,52 | 0,25% |
| 50-55 | 115,81 | 0,25% | 49,24 | 0,18% |
| 45-50 | 77,65 | 0,17% | 32,68 | 0,12% |
| 40-45 | 41,35 | 0,09% | 17,38 | 0,06% |
| 35-40 | 4,56 | 0,01% | 1,93 | 0,01% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #6 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 1 682,09 | 79,27% | 2 454,73 | 85,13% |
| 90-95 | 190,66 | 8,98% | 208,42 | 7,23% |
| 85-90 | 89,49 | 4,22% | 89,12 | 3,09% |
| 80-85 | 53,84 | 2,54% | 48,36 | 1,68% |
| 75-80 | 35,86 | 1,69% | 30,04 | 1,04% |
| 70-75 | 24,97 | 1,18% | 19,85 | 0,69% |
| 65-70 | 17,44 | 0,82% | 13,18 | 0,46% |
| 60-65 | 12,14 | 0,57% | 9,04 | 0,31% |
| 55-60 | 8,21 | 0,39% | 6,03 | 0,21% |
| 50-55 | 5,09 | 0,24% | 3,57 | 0,12% |
| 45-50 | 2,27 | 0,11% | 1,28 | 0,04% |
| 40-45 | 0,00 | 0,00% | 0,00 | 0,00% |
| 35-40 | 0,00 | 0,00% | 0,00 | 0,00% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #7 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 13 179,77 | 83,72% | 13 424,11 | 84,81% |
| 90-95 | 1 190,85 | 7,56% | 1 164,43 | 7,36% |
| 85-90 | 508,27 | 3,23% | 493,32 | 3,12% |
| 80-85 | 287,89 | 1,83% | 260,48 | 1,65% |
| 75-80 | 186,43 | 1,18% | 161,21 | 1,02% |
| 70-75 | 129,66 | 0,82% | 110,30 | 0,70% |
| 65-70 | 92,44 | 0,59% | 78,29 | 0,49% |
| 60-65 | 66,46 | 0,42% | 56,47 | 0,36% |
| 55-60 | 47,86 | 0,30% | 39,73 | 0,25% |
| 50-55 | 31,81 | 0,20% | 25,50 | 0,16% |
| 45-50 | 17,72 | 0,11% | 13,60 | 0,09% |
| 40-45 | 2,74 | 0,02% | 1,64 | 0,01% |
| 35-40 | 0,00 | 0,00% | 0,00 | 0,00% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #8 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 7 461,80 | 85,40% | 6 456,36 | 81,16% |
| 90-95 | 499,95 | 5,72% | 637,47 | 8,01% |
| 85-90 | 259,31 | 2,97% | 304,14 | 3,82% |
| 80-85 | 156,81 | 1,79% | 181,59 | 2,28% |
| 75-80 | 107,26 | 1,23% | 115,41 | 1,45% |
| 70-75 | 76,78 | 0,88% | 79,92 | 1,00% |
| 65-70 | 56,04 | 0,64% | 58,62 | 0,74% |
| 60-65 | 41,58 | 0,48% | 43,64 | 0,55% |
| 55-60 | 30,90 | 0,35% | 31,94 | 0,40% |
| 50-55 | 22,79 | 0,26% | 22,95 | 0,29% |
| 45-50 | 15,26 | 0,17% | 15,27 | 0,19% |
| 40-45 | 8,43 | 0,10% | 7,39 | 0,09% |
| 35-40 | 0,74 | 0,01% | 0,46 | 0,01% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #9 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 19 162,54 | 84,10% | 21 041,81 | 84,78% |
| 90-95 | 1 610,33 | 7,07% | 1 790,00 | 7,21% |
| 85-90 | 711,55 | 3,12% | 769,43 | 3,10% |
| 80-85 | 402,86 | 1,77% | 410,57 | 1,65% |
| 75-80 | 267,49 | 1,17% | 253,47 | 1,02% |
| 70-75 | 192,29 | 0,84% | 176,81 | 0,71% |
| 65-70 | 139,99 | 0,61% | 126,28 | 0,51% |
| 60-65 | 103,75 | 0,46% | 90,90 | 0,37% |
| 55-60 | 76,31 | 0,33% | 66,06 | 0,27% |
| 50-55 | 55,14 | 0,24% | 46,42 | 0,19% |
| 45-50 | 37,37 | 0,16% | 30,05 | 0,12% |
| 40-45 | 22,54 | 0,10% | 16,05 | 0,06% |
| 35-40 | 4,59 | 0,02% | 2,32 | 0,01% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #10 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 1 613,52 | 84,72% | 890,91 | 79,75% |
| 90-95 | 138,00 | 7,25% | 109,72 | 9,82% |
| 85-90 | 59,34 | 3,12% | 41,82 | 3,74% |
| 80-85 | 34,88 | 1,83% | 26,49 | 2,37% |
| 75-80 | 22,52 | 1,18% | 18,29 | 1,64% |
| 70-75 | 14,94 | 0,78% | 12,35 | 1,11% |
| 65-70 | 9,80 | 0,51% | 8,37 | 0,75% |
| 60-65 | 6,48 | 0,34% | 5,50 | 0,49% |
| 55-60 | 3,89 | 0,20% | 3,02 | 0,27% |
| 50-55 | 1,11 | 0,06% | 0,72 | 0,06% |
| 45-50 | 0,00 | 0,00% | 0,00 | 0,00% |
| 40-45 | 0,00 | 0,00% | 0,00 | 0,00% |
| 35-40 | 0,00 | 0,00% | 0,00 | 0,00% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #11 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 4 706,64 | 70,97% | 3 130,18 | 76,63% |
| 90-95 | 767,05 | 11,57% | 396,39 | 9,70% |
| 85-90 | 430,46 | 6,49% | 211,15 | 5,17% |
| 80-85 | 274,08 | 4,13% | 132,54 | 3,24% |
| 75-80 | 186,46 | 2,81% | 88,34 | 2,16% |
| 70-75 | 125,40 | 1,89% | 59,30 | 1,45% |
| 65-70 | 81,36 | 1,23% | 38,97 | 0,95% |
| 60-65 | 47,14 | 0,71% | 22,38 | 0,55% |
| 55-60 | 12,96 | 0,20% | 5,36 | 0,13% |
| 50-55 | 0,00 | 0,00% | 0,00 | 0,00% |
| 45-50 | 0,00 | 0,00% | 0,00 | 0,00% |
| 40-45 | 0,00 | 0,00% | 0,00 | 0,00% |
| 35-40 | 0,00 | 0,00% | 0,00 | 0,00% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #12 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 522,22 | 88,38% | 184,16 | 72,27% |
| 90-95 | 22,75 | 3,85% | 26,26 | 10,31% |
| 85-90 | 14,48 | 2,45% | 14,10 | 5,53% |
| 80-85 | 9,44 | 1,60% | 8,74 | 3,43% |
| 75-80 | 6,44 | 1,09% | 6,28 | 2,46% |
| 70-75 | 4,86 | 0,82% | 4,88 | 1,92% |
| 65-70 | 3,48 | 0,59% | 3,54 | 1,39% |
| 60-65 | 2,61 | 0,44% | 2,73 | 1,07% |
| 55-60 | 1,94 | 0,33% | 1,91 | 0,75% |
| 50-55 | 1,38 | 0,23% | 1,27 | 0,50% |
| 45-50 | 0,90 | 0,15% | 0,74 | 0,29% |
| 40-45 | 0,36 | 0,06% | 0,22 | 0,08% |
| 35-40 | 0,00 | 0,00% | 0,00 | 0,00% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #13 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 1 373,85 | 84,56% | 442,78 | 82,38% |
| 90-95 | 94,40 | 5,81% | 35,33 | 6,57% |
| 85-90 | 52,23 | 3,21% | 19,69 | 3,66% |
| 80-85 | 34,31 | 2,11% | 12,95 | 2,41% |
| 75-80 | 23,72 | 1,46% | 9,10 | 1,69% |
| 70-75 | 16,88 | 1,04% | 6,55 | 1,22% |
| 65-70 | 12,14 | 0,75% | 4,69 | 0,87% |
| 60-65 | 8,51 | 0,52% | 3,28 | 0,61% |
| 55-60 | 5,55 | 0,34% | 2,12 | 0,40% |
| 50-55 | 2,97 | 0,18% | 1,00 | 0,19% |
| 45-50 | 0,16 | 0,01% | 0,03 | 0,01% |
| 40-45 | 0,00 | 0,00% | 0,00 | 0,00% |
| 35-40 | 0,00 | 0,00% | 0,00 | 0,00% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #14 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 1 461,32 | 83,81% | 1 541,32 | 88,22% |
| 90-95 | 105,87 | 6,07% | 83,88 | 4,80% |
| 85-90 | 53,40 | 3,06% | 40,81 | 2,34% |
| 80-85 | 35,36 | 2,03% | 24,97 | 1,43% |
| 75-80 | 25,94 | 1,49% | 17,09 | 0,98% |
| 70-75 | 19,32 | 1,11% | 12,54 | 0,72% |
| 65-70 | 14,24 | 0,82% | 9,36 | 0,54% |
| 60-65 | 10,76 | 0,62% | 6,81 | 0,39% |
| 55-60 | 8,00 | 0,46% | 4,97 | 0,28% |
| 50-55 | 5,69 | 0,33% | 3,39 | 0,19% |
| 45-50 | 3,38 | 0,19% | 1,82 | 0,10% |
| 40-45 | 0,35 | 0,02% | 0,13 | 0,01% |
| 35-40 | 0,00 | 0,00% | 0,00 | 0,00% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #15 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 2 413,17 | 89,28% | 2 977,61 | 91,71% |
| 90-95 | 116,21 | 4,30% | 114,58 | 3,53% |
| 85-90 | 65,84 | 2,44% | 60,59 | 1,87% |
| 80-85 | 42,84 | 1,58% | 39,01 | 1,20% |
| 75-80 | 28,28 | 1,05% | 25,41 | 0,78% |
| 70-75 | 19,01 | 0,70% | 16,42 | 0,51% |
| 65-70 | 12,21 | 0,45% | 9,54 | 0,29% |
| 60-65 | 5,49 | 0,20% | 3,52 | 0,11% |
| 55-60 | 0,00 | 0,00% | 0,00 | 0,00% |
| 50-55 | 0,00 | 0,00% | 0,00 | 0,00% |
| 45-50 | 0,00 | 0,00% | 0,00 | 0,00% |
| 40-45 | 0,00 | 0,00% | 0,00 | 0,00% |
| 35-40 | 0,00 | 0,00% | 0,00 | 0,00% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #16 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 20 324,70 | 78,70% | 38 376,81 | 82,99% |
| 90-95 | 2 144,29 | 8,30% | 3 307,40 | 7,15% |
| 85-90 | 1 133,54 | 4,39% | 1 657,39 | 3,58% |
| 80-85 | 737,65 | 2,86% | 1 026,32 | 2,22% |
| 75-80 | 511,24 | 1,98% | 695,70 | 1,50% |
| 70-75 | 363,09 | 1,41% | 486,72 | 1,05% |
| 65-70 | 276,64 | 1,07% | 334,80 | 0,72% |
| 60-65 | 188,03 | 0,73% | 214,05 | 0,46% |
| 55-60 | 116,09 | 0,45% | 116,28 | 0,25% |
| 50-55 | 31,81 | 0,12% | 29,21 | 0,06% |
| 45-50 | 0,00 | 0,00% | 0,00 | 0,00% |
| 40-45 | 0,00 | 0,00% | 0,00 | 0,00% |
| 35-40 | 0,00 | 0,00% | 0,00 | 0,00% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #17 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 372 521,59 | 93,19% | 83 681,09 | 92,64% |
| 90-95 | 14 321,67 | 3,58% | 3 554,44 | 3,93% |
| 85-90 | 5 001,02 | 1,25% | 1 257,95 | 1,39% |
| 80-85 | 2 693,62 | 0,67% | 660,94 | 0,73% |
| 75-80 | 1 687,50 | 0,42% | 406,61 | 0,45% |
| 70-75 | 1 154,96 | 0,29% | 267,98 | 0,30% |
| 65-70 | 820,81 | 0,21% | 184,34 | 0,20% |
| 60-65 | 587,35 | 0,15% | 127,14 | 0,14% |
| 55-60 | 412,30 | 0,10% | 86,04 | 0,10% |
| 50-55 | 280,72 | 0,07% | 55,76 | 0,06% |
| 45-50 | 180,30 | 0,05% | 32,63 | 0,04% |
| 40-45 | 89,51 | 0,02% | 13,81 | 0,02% |
| 35-40 | 1,84 | 0,00% | 0,20 | 0,00% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #18 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 41 752,89 | 90,36% | 93 440,06 | 83,67% |
| 90-95 | 2 202,03 | 4,77% | 9 234,57 | 8,27% |
| 85-90 | 1 130,13 | 2,45% | 4 622,94 | 4,14% |
| 80-85 | 631,16 | 1,37% | 2 577,04 | 2,31% |
| 75-80 | 343,21 | 0,74% | 1 346,10 | 1,21% |
| 70-75 | 147,78 | 0,32% | 458,01 | 0,41% |
| 65-70 | 0,00 | 0,00% | 0,00 | 0,00% |
| 60-65 | 0,00 | 0,00% | 0,00 | 0,00% |
| 55-60 | 0,00 | 0,00% | 0,00 | 0,00% |
| 50-55 | 0,00 | 0,00% | 0,00 | 0,00% |
| 45-50 | 0,00 | 0,00% | 0,00 | 0,00% |
| 40-45 | 0,00 | 0,00% | 0,00 | 0,00% |
| 35-40 | 0,00 | 0,00% | 0,00 | 0,00% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

| BANQUE DE DONNÉES #19 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 308 417,79 | 91,12% | 277 532,54 | 90,69% |
| 90-95 | 13 464,18 | 3,98% | 15 198,36 | 4,97% |
| 85-90 | 6 507,69 | 1,92% | 6 121,84 | 2,00% |
| 80-85 | 3 887,91 | 1,15% | 3 182,23 | 1,04% |
| 75-80 | 2 543,28 | 0,75% | 1 844,88 | 0,60% |
| 70-75 | 1 682,98 | 0,50% | 1 096,01 | 0,36% |
| 65-70 | 1 078,20 | 0,32% | 631,85 | 0,21% |
| 60-65 | 639,49 | 0,19% | 318,09 | 0,10% |
| 55-60 | 259,37 | 0,08% | 95,23 | 0,03% |
| 50-55 | 0,00 | 0,00% | 0,00 | 0,00% |
| 45-50 | 0,00 | 0,00% | 0,00 | 0,00% |
| 40-45 | 0,00 | 0,00% | 0,00 | 0,00% |
| 35-40 | 0,00 | 0,00% | 0,00 | 0,00% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

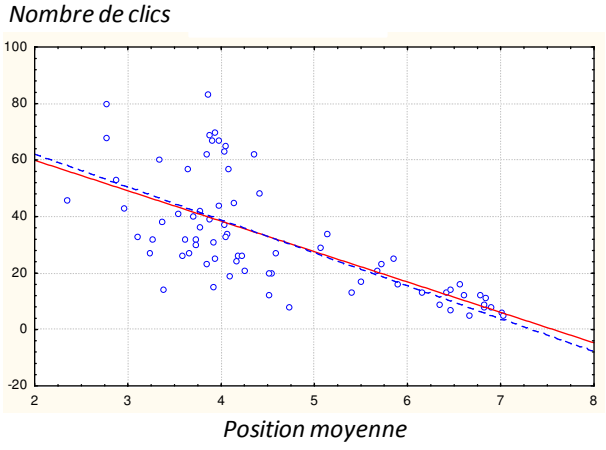
| BANQUE DE DONNÉES #20 | clics | | coûts | |
|---|---|---|---|---|
| groupe de mots-clés (centile) | somme des clics moyens par jour (clics) | proportion du total (%) | somme des coûts moyens par jour ($) | proportion du total (%) |
| 95-100 | 35 954,33 | 88,71% | 28 367,93 | 89,63% |
| 90-95 | 2 257,82 | 5,57% | 1 642,35 | 5,19% |
| 85-90 | 897,96 | 2,22% | 649,66 | 2,05% |
| 80-85 | 487,56 | 1,20% | 353,57 | 1,12% |
| 75-80 | 313,34 | 0,77% | 215,66 | 0,68% |
| 70-75 | 214,53 | 0,53% | 144,18 | 0,46% |
| 65-70 | 150,29 | 0,37% | 101,39 | 0,32% |
| 60-65 | 105,41 | 0,26% | 72,23 | 0,23% |
| 55-60 | 72,56 | 0,18% | 50,18 | 0,16% |
| 50-55 | 46,86 | 0,12% | 32,88 | 0,10% |
| 45-50 | 26,02 | 0,06% | 18,12 | 0,06% |
| 40-45 | 4,83 | 0,01% | 2,84 | 0,01% |
| 35-40 | 0,00 | 0,00% | 0,00 | 0,00% |
| 30-35 | 0,00 | 0,00% | 0,00 | 0,00% |
| 25-30 | 0,00 | 0,00% | 0,00 | 0,00% |
| 20-25 | 0,00 | 0,00% | 0,00 | 0,00% |
| 15-20 | 0,00 | 0,00% | 0,00 | 0,00% |
| 10-15 | 0,00 | 0,00% | 0,00 | 0,00% |
| 5-10 | 0,00 | 0,00% | 0,00 | 0,00% |
| 0-5 | 0,00 | 0,00% | 0,00 | 0,00% |

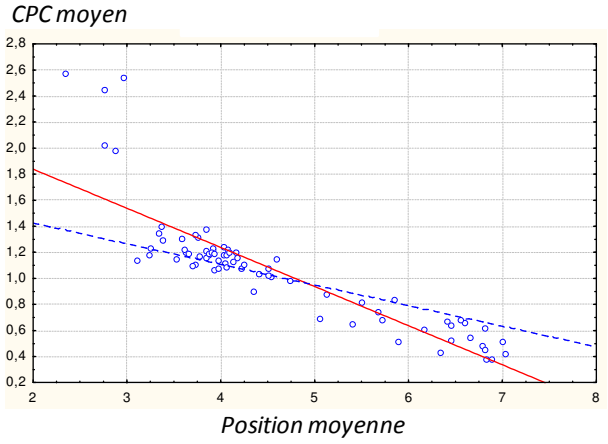
# ANNEXE 4 – Exemples de prédictions génériques linéaires et comparaison avec les régressions obtenues

Cette annexe présente six exemples de mots-clés pour lesquels nous avons utilisé des fonctions génériques de prédiction linéaires dans le but de prédire le comportement des clics et des CPC moyens. Dans chacun des graphiques, la fonction de prédiction générique calculée selon notre méthode (ligne bleue pointillée) est comparée à la fonction de régression qui s'ajuste aux observations (ligne pleine rouge).

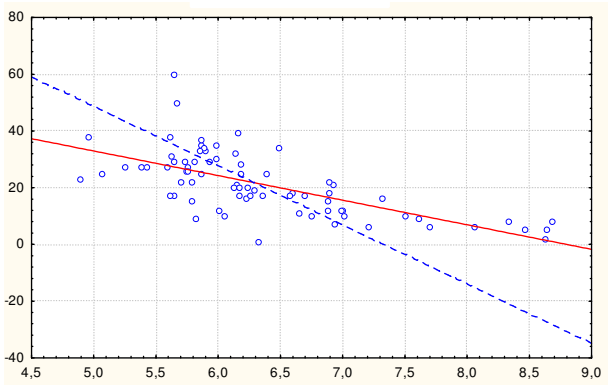
*Remarque #1* : Afin d'éviter les divisions par zéro, les graphiques de CPC moyen en fonction de la position moyenne n'illustrent pas les observations où il n'y a eu aucun clic (CPC moyen = coût/nombre de clics).

*Remarque #2* : Nous sommes conscients que les régressions présentées ne sont pas nécessairement toutes acceptables d'un point de vue statistique. Certaines d'entre elles ne respectent pas les conditions de normalité et d'indépendance des résidus. Cependant, nous les présentons tout de même, afin d'offrir une base de comparaison pour la fonction générique. Il est intéressant de comparer la fonction générique à la fonction obtenue par régression, car cette dernière minimise la somme du carré des erreurs sur les observations.

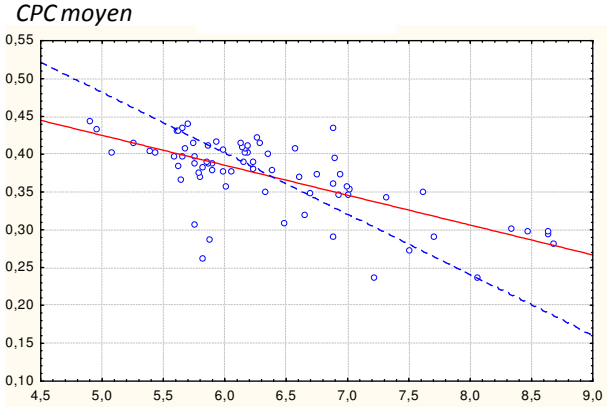


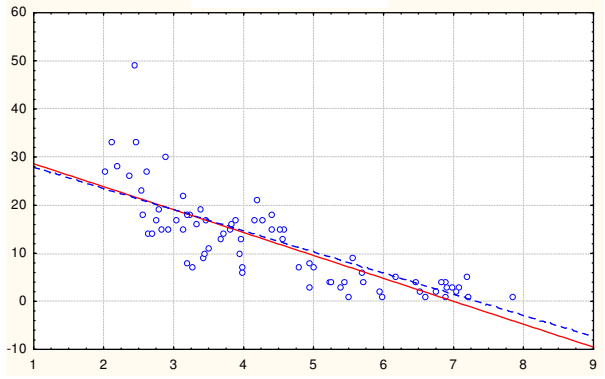
*Nombre de clics*



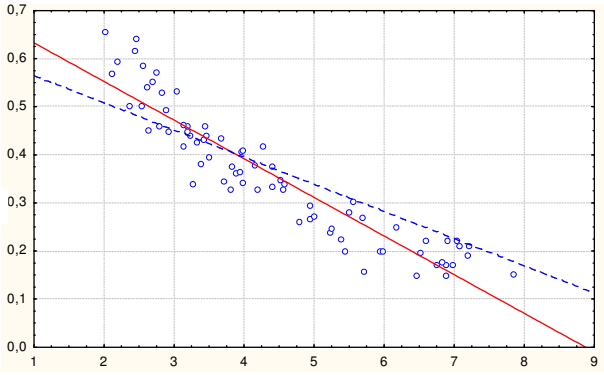

*CPC moyen*



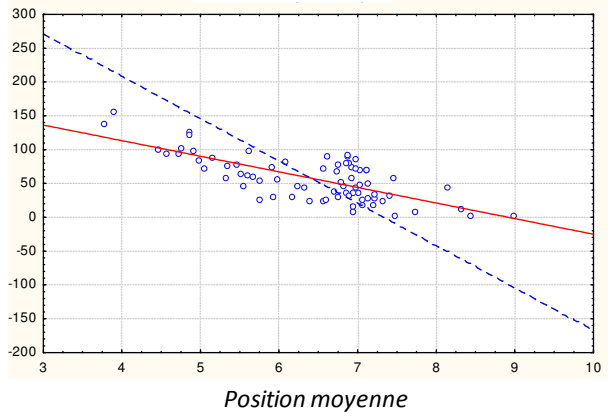

*Nombre de clics*



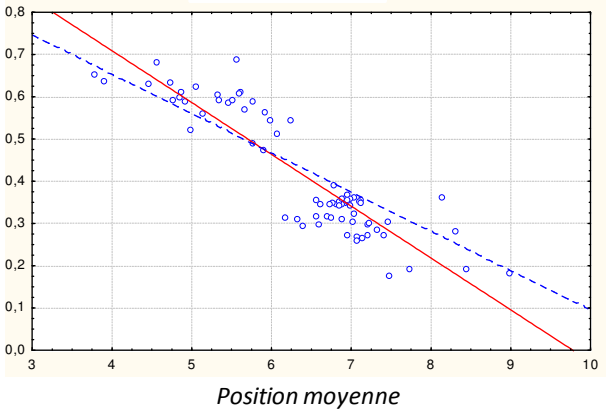

*CPC moyen*



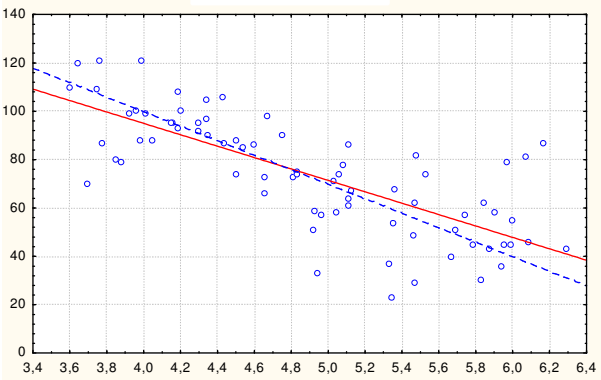

*Nombre de clics*




*CPC moyen*

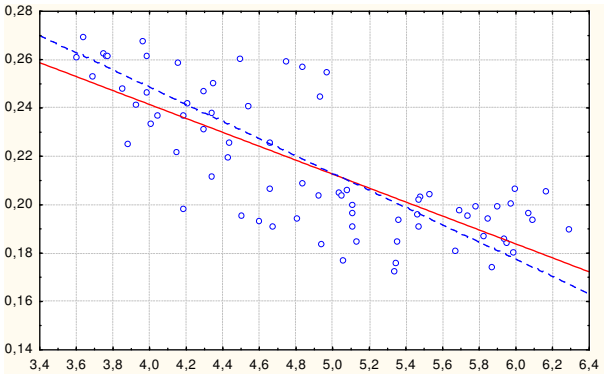


*Nombre de clics*



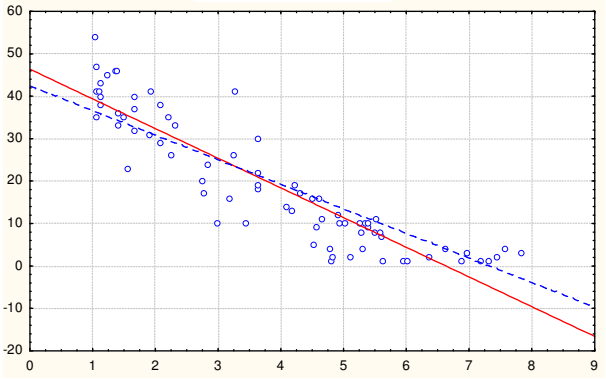

*Position moyenne*

*CPC moyen*



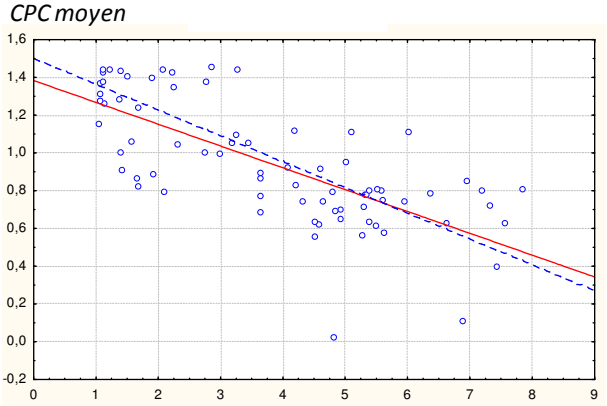

*Position moyenne*

*Nombre de clics*




*Position moyenne*

*CPC moyen*




*Position moyenne*

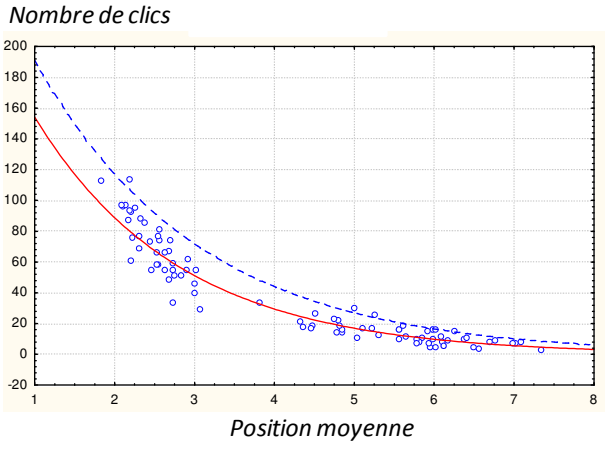
# ANNEXE 5 – Exemples de prédictions génériques exponentielles et comparaison avec les régressions obtenues

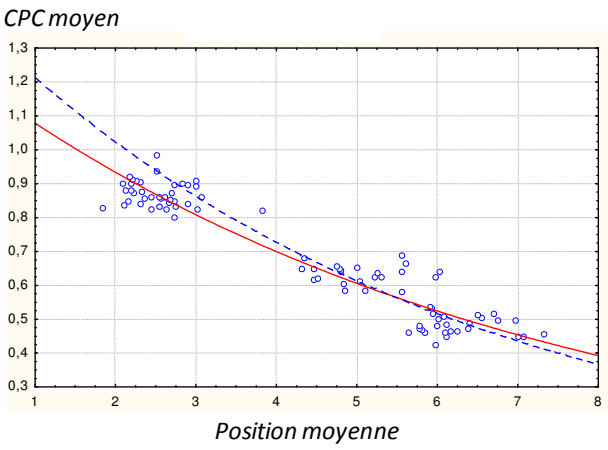
Cette annexe présente six exemples de mots-clés pour lesquels nous avons utilisé des fonctions génériques de prédiction exponentielles dans le but de prédire le comportement des clics et des CPC moyens. Dans chacun des graphiques, la fonction de prédiction générique calculée selon notre méthode (ligne bleue pointillée) est comparée à la fonction de régression qui s'ajuste aux observations (ligne pleine rouge).

*Remarque #1* : Afin d'éviter les divisions par zéro, les graphiques de CPC moyen en fonction de la position moyenne n'illustrent pas les observations où il n'y a eu aucun clic (CPC moyen = coût/nombre de clics).

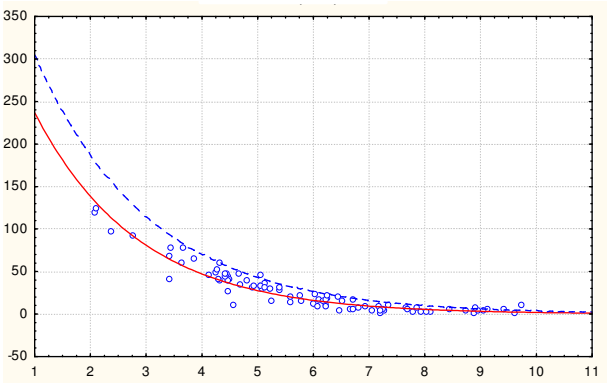
*Remarque #2* : Nous sommes conscients que les régressions présentées ne sont pas nécessairement toutes acceptables d'un point de vue statistique. Certaines d'entre elles ne respectent pas les conditions de normalité et d'indépendance des résidus. Cependant, nous les présentons tout de même, afin d'offrir une base de comparaison pour la fonction générique. Il est intéressant de comparer la fonction générique à la fonction obtenue par régression, car cette dernière minimise la somme du carré des erreurs sur les observations.

*Remarque #3* : Les régressions exponentielles ont été obtenues en effectuant des régressions linéaires sur les valeurs de $\ln(\text{nombre de clics})$ et $\ln(\text{CPC moyen})$ en fonction de la position moyenne.

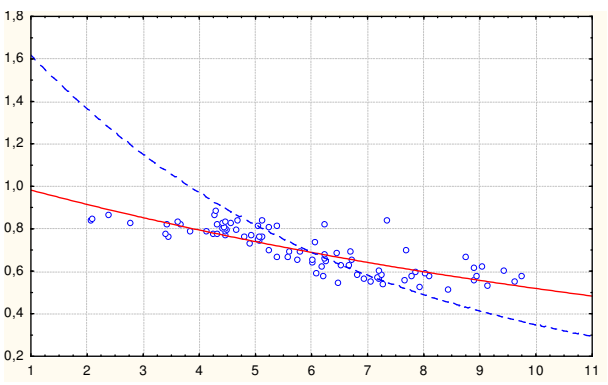


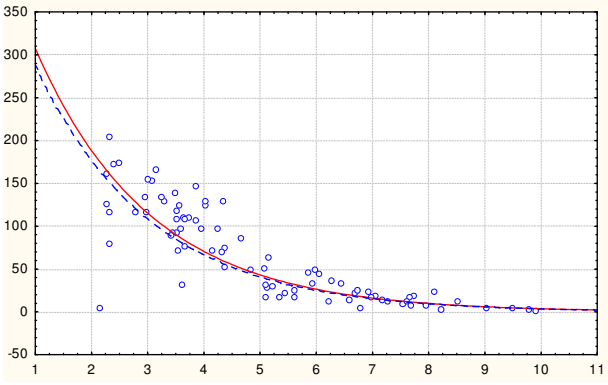
*Nombre de clics*



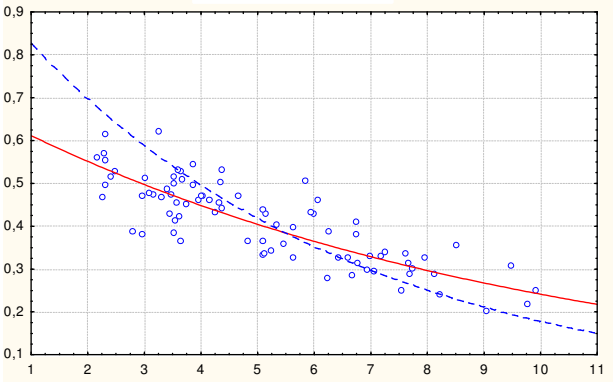

*CPC moyen*



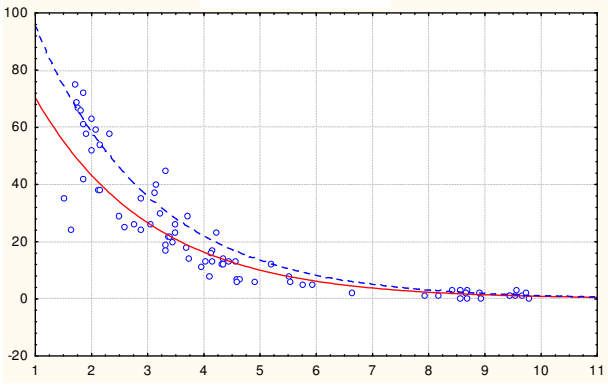

*Nombre de clics*



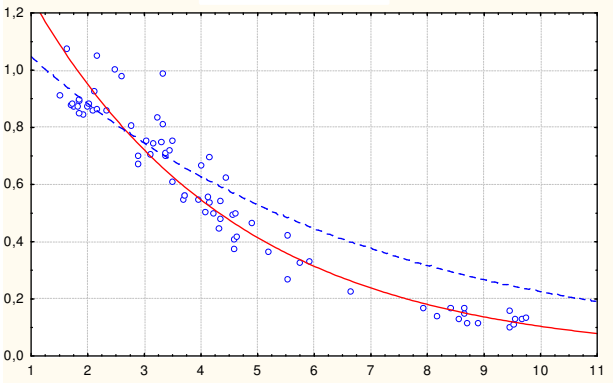

*CPC moyen*



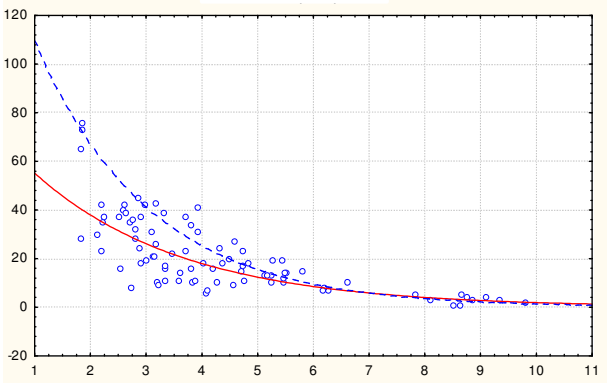

*Nombre de clics*




*CPC moyen*

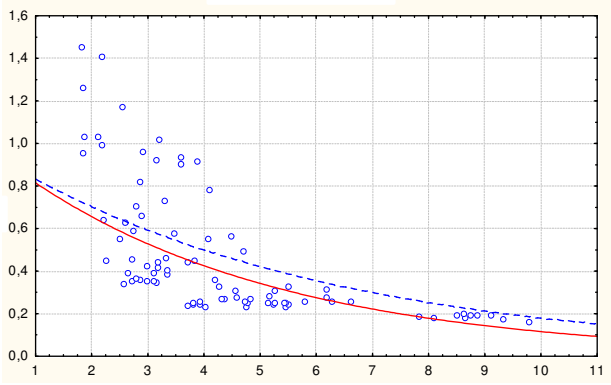


*Nombre de clics*



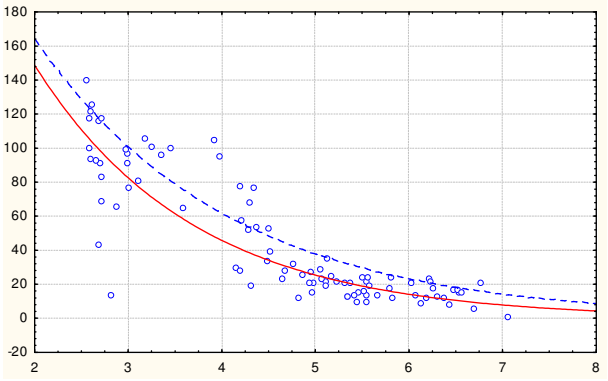

*Position moyenne*

*CPC moyen*



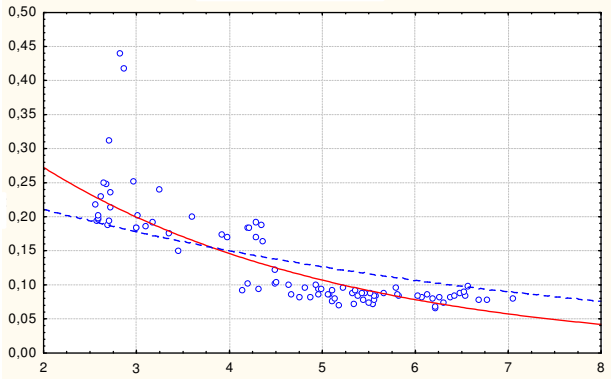

*Position moyenne*

*Nombre de clics*




*Position moyenne*

*CPC moyen*




*Position moyenne*