HEC Montréal

L'utilisation de données de sources multiples dans le développement de modèles statistiques : application à l'analyse de la performance de ventes en épargne dans une entreprise du secteur bancaire

par
Sarah Legendre Bilodeau

Sciences de la gestion
(Intelligence d'affaires)

Mémoire présenté en vue de l'obtention du grade de maîtrise ès sciences (M.Sc)

Décembre, 2014

Résumé

Les entreprises du secteur bancaire détiennent une importante quantité d'informations dans des bases de données internes. Ces bases de données sont utilisées par différents secteurs et sont exploitées afin de guider les gestionnaires dans leurs décisions. Or, des données provenant d'autres sources, moins fréquemment utilisées dans le secteur bancaire, pourraient permettre d'améliorer le niveau de connaissance du contexte d'affaires de l'entreprise. Ces données peuvent être internes à l'entreprise ou externes et peuvent constituer une valeur ajoutée non négligeable pour l'organisation.

L'objectif premier de ce projet de recherche est de vérifier si l'ajout de données provenant de multiples sources aux données des entrepôts de données d'une entreprise du secteur bancaire permet d'améliorer la performance de modèles statistiques. Un cas d'application sera utilisé dans le cadre de cette recherche. Plus précisément, l'intégration de données de fréquentation du site web de l'entreprise, en plus de considérer l'ajout de données provenant d'organismes de santé publique, de Statistique Canada, d'un outil de prestation de service et d'activités de sollicitation et de marketing direct sera considérée.

Afin de répondre adéquatement à cet objectif de recherche, un processus d'intégration de données est primordial. Dans le cadre de ce processus, une méthode d'appariement des données de fréquentation de sites web, lorsqu'elles sont extraites d'un outil d'analyse d'audience Internet, aux données internes de l'entreprise est développée. Afin de comparer l'utilisation de données d'entrepôts internes de l'entreprise avec l'utilisation de sources multiples de données, un modèle linéaire mixte dans un contexte de mesures répétées est considéré. Ensuite, une méthode de sélection de variables permettant de tenir compte du contexte de mesures répétées est développée et est utilisée afin de développer un modèle explicatif. Cette méthode peut être utilisée avec un grand nombre de variables et d'observations. Dans le contexte du cas d'application étudié, la méthode de sélection

de variables s'est avérée performante et l'ajout de données de sources multiples a permis, de façon générale, le développement de modèles plus précis. Des éléments pertinents dans le cadre du développement d'un modèle prédictif sont discutés.

Mots-clés : multiples sources de données, modèles linéaires mixtes, mesures répétées, données longitudinales, intégration des données, sélection de modèles, données web, données externes, analyse de la performance.

Table des matières

Liste des tableaux

Liste des figures

À mes enfants, Léandre et Rose, à qui
j'espère avoir donné le goût des études

Remerciements

Tout d'abord, je souhaite remercier mon directeur de recherche, Marc Fredette, qui a su guider mon travail de façon efficace au cours des dernières années. Il s'est intéressé au projet de recherche qui me tenait à cœur et m'a encouragé dans mon ambition d'utiliser des données moins conventionnelles. Il s'est adapté à la réalité qui était la mienne et je lui en suis très reconnaissante.

Je remercie également mes collègues de travail qui m'ont encouragé tout au long de mes études. Un merci tout spécial à Julie grâce à qui j'ai pu utiliser des données de fréquentation de sites web et qui a toujours été disponible pour répondre à mes nombreuses questions. Merci également à David qui, avec ses petites suggestions qui arrivaient au bon moment, me permettait d'aller plus loin et de résoudre des problèmes. Merci à Josée pour son efficacité, les données de sollicitation et les réponses à mes questions. Merci à Elahdji pour ses citations appropriées, pour son soutien et pour le plaisir de discuter. Merci à Virginie, ma collègue et amie, qui prend toujours le temps de s'informer sur les avancées de mon travail et qui me soutient dans les moments difficiles. Merci à Martine pour son soutien efficace et son sens de l'humour. Merci à mes patrons, actuel et passés, Christian, Sébastien, Jean-François et Claudine, d'avoir cru en mes capacités et de m'avoir soutenu pendant ces quelques années de maîtrise. Ils ont fait preuve d'ouverture d'esprit et m'ont donné beaucoup de liberté et de flexibilité afin que je sois en mesure de réaliser ce projet important à mes yeux.

À mon amie de longue date, Paskale, avec qui j'ai parcouru beaucoup de chemin, merci d'être présente dans ma vie. Que ce soit au cours de mes études secondaires, pendant nos années d'études en musique au Conservatoire, dans ma vie de famille, dans les nombreux concerts que nous avons faits ensemble, elle a toujours été là, de près ou de loin, avec les mots appropriés et une grande implication.

Finalement, merci à André, mon conjoint depuis quinze ans et père de mes enfants, qui a cru en moi et m'a soutenu pendant toutes ces années. Merci de m'avoir aidé à persévérer tout au long de mes études et de reconnaître de si belle façon tous mes efforts. Merci également à mes enfants, pour qui ces années d'études n'ont pas toujours été faciles, mais qui ont eu la maturité de comprendre les efforts investis.

Introduction

Les entreprises du secteur bancaire détiennent une importante quantité d'informations dans des bases de données internes. Ces bases de données sont utilisées par différents secteurs et sont exploitées afin de guider les gestionnaires dans leurs décisions. Ces décisions peuvent être de natures très variées : clientèle, géographie, distribution, ventes, développement des affaires, développement de produits. Les bases de données sont nombreuses, souvent étendues dans l'organisation et de formats variés, ce qui rend difficile l'utilisation simultanée de plusieurs sources de données dans un projet d'intelligence d'affaires. De plus, l'accès aux données est souvent limité par des mesures de sécurité importantes et multiples. Ces quelques raisons peuvent expliquer en partie la pratique encore répandue de l'utilisation de sources limitées ou uniques de données dans la construction de modèles statistiques explicatifs et prédictifs. Dans un autre ordre d'idées, le contexte actuel de rationalisation et de souci d'augmentation de la productivité ouvre très peu la porte à l'investissement important en temps lors de l'étape de consolidation des données nécessaires qui serviront à effectuer un mandat, souvent requis à l'intérieur de courts délais. S'ajoute à ces éléments l'importance d'avoir des outils technologiques performants pour le traitement efficace de données volumineuses, sans lesquels l'intégration de données variées devient pratiquement impossible. Or, les sources multiples de données dans une entreprise constituent une importante valeur, qui peut être mise à profit de meilleure façon.

Dans ce contexte, la présente étude vise à vérifier si l'intégration de données de sources multiples, dans certains cas externes à l'entreprise, améliore de façon non négligeable la performance de modèles explicatifs, et éventuellement prédictifs, par rapport à l'utilisation d'une source unique interne à l'entreprise. Les données provenant de la fréquentation d'un site web constituent un exemple de données de source externe qui sera considérée. Un cas d'application particulier est utilisé dans cette étude et les conclusions de cette étude seront relatives à ce cas d'application. Les données utilisées

pour le cas d'application proviennent d'une entreprise du secteur bancaire ayant ses activités notamment au Québec et en Ontario.

Chapitre 1 : Revue de littérature

Les publications scientifiques discutant de l'utilisation de données de sources multiples ou externes pour la construction de modèles statistiques explicatifs et prédictifs ayant trait à l'analyse de la performance de ventes de produits financiers sont rares. En effet, il s'agit d'une pratique récente qui se développe de pair avec la montée en popularité de l'expression « Big Data » (ou données volumineuses) depuis quelques années. Il importe donc d'élargir le champ de recherche à d'autres secteurs d'activités, où les types de données utilisées et les défis auxquels sont confrontés les chercheurs peuvent être similaires en plusieurs points à ce qui est vécu dans le secteur financier. Cette section de la revue de littérature abordera le sujet de façon assez large et tentera de traiter de différents points d'intérêts qui seront abordés dans cette recherche. Dès lors, une section sera consacrée à l'utilisation de données de sources multiples dans le domaine de la santé publique, étant donné son importance dans la littérature.

Les premières publications scientifiques traitant de l'utilisation de données de sources variées, peu importe les domaines d'applications, sont apparues au cours des années 1990, ce qui n'est pas étonnant considérant les ressources logicielles importantes requises dans de tels cas. Des volumes importants de données sont disponibles au sein des organisations, dans une variété de formats plus ou moins conventionnels (Zeng et Tang, 2008). Comme mentionné par Zeng et Tang (2008), les entreprises ont besoin de mieux utiliser leurs ressources d'informations afin d'être plus efficientes et efficaces, surtout dans le contexte d'économie globale actuel. Ainsi, ces auteurs ont entrepris de faire une revue des usages de sources de données multiples dans les entreprises pour orienter la prise de décision et la planification de leurs activités. Même si ces recherches ne couvrent pas le volet de l'utilisation des données de sources multiples dans un contexte statistique ou d'analyse à proprement parler, les éléments ayant trait à

l'intégration sont pertinents, puisqu'ils touchent aux avantages, difficultés et enjeux qui peuvent être vécus dans le cas d'application qui nous intéresse. Ainsi, Zeng et Tang (2008) ont identifié les recherches courantes et les expériences pratiques relativement à l'utilisation de sources de données multiples afin de supporter la mesure de la performance, la planification stratégique et les processus d'affaires interorganisations. Ils mentionnent que le but principal de leur publication est de dresser les premières étapes pour le développement d'une méthodologie d'intégration de données de sources multiples.

1.1 Intégration des données

Dans leur article, Zeng et Tang (2008) font de l'intégration des données un point central de l'utilisation de données de sources variées. Ils mentionnent que l'intégration des données apporte une solution au fait que beaucoup d'informations sont générées en silos. Ils ajoutent également qu'elles sont difficilement accessibles au sein de leur organisation et encore plus difficilement aux organismes connexes extérieurs à l'organisation. Ils définissent l'intégration des données (*Data Integration*) par un processus de standardisation dans la définition et la structure des données par l'utilisation d'un schéma conceptuel commun (Heimbigner et McLeod, 1985). En complément, l'intégration des données doit être compatible de façon constante et logique avec différents systèmes ou bases de données et pouvoir être utilisée à travers le temps et les utilisateurs (Martin, 1986, cité par Zeng et Tang, 2008).

L'objectif de l'intégration des données est de rendre disponibles des données provenant de sources multiples permettant de dégager l'information contribuant à la réalisation des objectifs des utilisateurs (AFT, 1997, cité par Zeng et Tang, 2008). L'identification d'un sous-ensemble optimal de sources disponibles, l'estimation des niveaux de bruits et de distorsions liés à la collecte de données, la précision et le format des données,

l'entreposage et la récupération des données, la performance liée à l'atteinte des objectifs des utilisateurs doivent être considérés lors d'une intégration adéquate des données (AFT, 1997, cité par Zeng et Tang, 2008).

Certains bénéfices liés à l'intégration peuvent être mis de l'avant, soient de limiter les efforts requis pour la collecte de données, de permettre une extraction d'informations qui serait autrement impossible (Subrahmanian et al., 1996, cité par Zeng et Tang, 2008) en prenant chaque source de façon isolée, d'améliorer l'information de gestion dans l'entreprise pour la prise de décisions et la planification stratégique et d'avoir une meilleure coordination opérationnelle entre les différents secteurs (Zeng et Tang, 2008). De plus, Zeng et Tang (2008) mentionnent que l'intégration des données peut avoir un impact positif en réduisant les coûts et en diminuant les éléments d'ambiguïté entre les différents secteurs de l'organisation. En opposition, ils ont également constaté que parce que plusieurs secteurs sont impliqués, l'intégration des données peut également augmenter les coûts en augmentant la taille et la complexité du problème de conception ou augmenter la difficulté à obtenir l'accord de toutes les parties concernées. Plus en détail, ils relèvent trois entraves majeures :

- la nécessité de faire des compromis en matière de satisfaction des besoins locaux en données et en informations;
- les délais de nature bureaucratique réduisant la flexibilité locale;
- la hausse des coûts dans la conception et la mise en place des systèmes d'information.

Le sujet de l'intégration des données est intéressant dans le contexte de cette étude, puisqu'il met en lumière plusieurs éléments qui peuvent aussi être constatés ou considérés dans le cas d'un travail isolé d'utilisation de sources multiples de données. En effet, l'utilisation de données de sources multiples dans le cas d'application étudié dans le cadre de ce projet de recherche peut être vue comme étant un effort non récurrent d'intégration des données, avec ses avantages, inconvénients et enjeux. De plus, il arrive

qu'une analyse non récurrente soit transformée en un projet de mise en production si l'utilité de l'analyse est confirmée et que les avantages liés à une utilisation récurrente sont certains. Dans de tels cas, l'intégration des données en bonne et due forme est essentielle en considérant les bénéfices identifiés précédemment.

1.2 Utilisation de données de sources multiples dans un contexte de santé publique

Plusieurs recherches portant sur l'utilisation de sources de données multiples sont effectuées dans un contexte de santé publique ou de recherche en épidémiologie. La concentration des recherches dans ce domaine d'expertise est probablement liée au fait que les différentes entités de santé sont relativement indépendantes et que la gestion des sources de données variées est très fréquente chez les chercheurs dans ce domaine. Ainsi, la nécessité de se questionner au sujet des méthodes à utiliser, des bénéfices et des enjeux liés à ce type de pratique est plus criante.

Dans un récent article, Bradley *et al.* (2010) discutent de l'utilisation de sources multiples de données afin de bonifier l'efficacité de la recherche en services de santé. Ainsi, les possibilités et avantages qu'offre ce type de pratique sont discutés, de même que les limitations et les défis qui y sont rattachés. De plus, ils proposent une méthodologie appropriée lorsque l'intégration de sources multiples de données est envisagée et ils présentent quelques pistes permettant de fusionner les différentes informations pour chaque identifiant.

Plusieurs éléments abordés par Bradley *et al.* (2010) sont pertinents dans notre contexte d'étude. Mentionnons tout d'abord le contexte actuel d'analyses multiples, effectuées par plusieurs chercheurs de départements ou organisations variés, à partir de sources de

données localisées et à portées limitées, qui ne sont pas mises en commun ou coordonnées. Ce type de pratique est observé également dans plusieurs entreprises privées, dont celle qui sert de cas d'application dans cette présente recherche. Effectivement, même s'il y a un souci de consolider l'information de l'entreprise au sein d'une base de données principale, plusieurs secteurs font le choix d'alimenter des bases de données plus spécialisées, selon leurs besoins. Cette pratique est très répandue dans les secteurs qui traitent des données moins conventionnelles, telles que des données de sites web ou de médias sociaux. Selon Bradley *et al.* (2010), les pratiques actuelles de travail sur des données en silo impliquent que des efforts majeurs doivent être déployés afin de combiner l'information provenant de différentes sources.

D'après Bradley *et al.* (2010), les sources de données peuvent être regroupées et catégorisées en fonction de l'agent majeur de contrôle de la donnée. Dans le cas de recherches en santé, ils identifient les catégories suivantes d'agents :

1. chercheurs individuels;
2. bases de données sous responsabilité gouvernementale;
3. partenariats public-privé favorisant l'utilisation et l'intégration de données produites et contrôlées par des organisations privées.

Les auteurs mentionnent que ces différents agents décident de la disponibilité de la donnée et de la façon dont elle peut être extraite, intégrée et utilisée par les chercheurs. Ces concepts peuvent aisément être adaptés au contexte de données d'une entreprise bancaire.

De plus, Bradley *et al.* (2010) donnent une procédure à suivre lors d'un projet impliquant l'utilisation de données de sources multiples. Ainsi, ils décrivent cinq étapes de base à suivre, qui peuvent être résumées ainsi :

1. identifier les sources de données qui pourront répondre à une question précise de recherche;

2. obtenir les droits d'utilisation;
3. sélectionner les variables qui seront utilisées lors de la fusion et nettoyer les données;
4. choisir la meilleure méthode de fusion et développer les algorithmes appropriés;
5. évaluer la qualité des liens entre les sources de données.

Ces étapes ne sont pas seulement pertinentes dans les cas de fusions de données dans le domaine de la santé publique, mais dans tous les cas où on peut répondre à une problématique à l'aide de l'exploitation de données de sources diverses.

Toujours selon les chercheurs, l'intégration de données de sources multiples implique plusieurs limitations et contraintes qui ne doivent pas être minimisées. Parmi ces limitations et contraintes, ils mentionnent les éléments suivants :

- ressources financières et temps requis élevés;
- haut degré de responsabilité relativement à la protection, l'entreposage et l'utilisation des données;
- portée de la population étudiée dans les données secondaires;
- habileté élevée requise portant sur l'extraction et l'imputation des données requises;
- expertises requises dans plusieurs domaines (sources de données, programmation, analyses statistiques, interprétation), impliquant la collaboration entre plusieurs types de spécialistes;
- pourvoyeurs multiples, moments de disponibilité des données variés, accessibilité, coût des données, contraintes légales et approbations.

Bradley *et al.* (2010) discutent ensuite des prérequis pour lier les données entre elles. Ils mentionnent qu'au moins un identifiant commun est requis dans les sources qui doivent être liées. Selon eux, ces identifiants peuvent être à un niveau très fin ou à un niveau

d'agrégation plus élevé. De plus, ils mentionnent que la qualité des identifiants peut parfois laisser à désirer, ce qui peut nécessiter l'ajout d'autres variables permettant de lier les données entre elles avec plus de certitude, et qu'il est important de se préoccuper de la présence de doublons dans les sources de données qui doivent être liées.

Deux types de méthodes de liaison de données sont mentionnés : l'appariement déterministe et l'appariement probabiliste. Selon les auteurs, ces types impliquent une série d'étapes à exécuter dans un ordre précis pour apparier les données dans le cas déterministe, ou une forte probabilité que deux enregistrements concernent un même identifiant dans le cas probabiliste. Des modèles mathématiques ont été développés afin de considérer des enregistrements appariés (Fellegi et Sunter, 1969, cité par Bradley *et al.*, 2010) et des investissements logiciels sont requis lorsque l'appariement probabiliste doit être utilisé.

Finalement, Bradley *et al.* (2010) ont émis plusieurs recommandations d'importance afin de permettre une meilleure accessibilité aux données par les chercheurs et pour améliorer les systèmes d'informations. Selon eux, la plupart de ces recommandations favorisent une approche systématique et centralisée, ce qui permet de maximiser les systèmes de données déjà en place et de combler les lacunes dans l'infrastructure existante. Ces recommandations de Bradley *et al.* (2010) se résument ainsi :

- en collaboration avec différents secteurs, convenir d'un plan d'amélioration des systèmes d'entreposage de données déjà en place;
- éliminer les entraves, en identifiant les pratiques qui contraignent l'utilisation des données de sources variées;
- mettre en place des standards inspirés des meilleurs pratiques en appariement de données, utilisation de données secondaires, qualité des données, confidentialité et entreposage de données;

- mettre à profit les bases de données sectorielles existantes, en appuyant les secteurs pour des développements centralisés de bases de données relationnelles;
- collecter et intégrer des données qualitatives en tant que source secondaire.

Alors que Bradley *et al.* (2010) ont surtout discuté des processus dans le cas d'intégration de données de sources multiples, d'autres chercheurs, toujours dans le domaine de la santé publique, s'attardent aux données de sources variées afin de répondre à des besoins bien précis. Ainsi, Molitor *et al.* (2009) mentionnent que l'utilisation de données de sources variées permet de déterminer différents biais qui seraient difficiles à détecter avec l'utilisation d'une seule source de données, lorsqu'une question de recherche est posée. Ils parlent donc plus de la pertinence de les utiliser. Selon eux, l'utilisation de données de sources multiples est utile pour deux raisons principales : bénéficier de l'augmentation de la puissance d'analyse que procure l'utilisation de données combinées et répondre à certaines incertitudes dues à des données manquantes.

Dans leur pratique d'utilisation de sources multiples, ils adoptent un modèle graphique bayésien, qui tient compte de la construction de sous-modèles locaux en fonction des différentes sources de données, qui sont ensuite intégrés en une seule analyse globale (Spiegelhalter, 1998, Richardson et Best, 2003, cités par Molinor *et al.*, 2009). Mentionnons que Molitor *et al.* (2009), dans leur cas d'application, complètent les informations obtenues à partir de données administratives et de données d'enquête par l'ajout de données agrégées permettant d'ajouter des précisions. Ils mentionnent que ces données agrégées sont ajoutées au niveau de la région ou du code postal.

Molitor *et al.* (2009) identifient plusieurs avantages d'un modèle unifié :

- l'estimation des paramètres est effectuée simultanément, peu importe la source de données, dans un seul modèle;
- l'intégration des différentes sources de données dans le modèle est faite de façon cohérente;

- l'explication du modèle unifié peut se faire de façon relativement simple à comprendre à l'aide de la combinaison des sous-modèles.

Dans leur étude, Molitor *et al.* (2009) ont constaté certains éléments concernant la pertinence d'utiliser des données de sources variées. Ainsi, à l'aide de simulations, ils ont montré que les modèles d'imputation donnent de bons résultats lorsque les variables explicatives sont fortement corrélées avec les variables étudiées dans chaque sous-modèle du graphique. De plus, ils mentionnent que l'application de leur modèle dans un contexte d'étude épidémiologique a permis d'améliorer l'estimation de la variable dépendante lorsque des données de sources multiples étaient utilisées, en opposition à l'utilisation de données d'une source unique. Toutefois, ils ont également démontré que lorsque les informations utilisées provenant des différentes sources de données sont peu liées entre elles, la combinaison des différentes sources ne donne pas beaucoup d'avantages par rapport à une source unique de données.

Molitor *et al.* (2009) ont mentionné un désavantage principal au sujet de l'utilisation d'un modèle graphique Bayesien, soit les importantes ressources logicielles requises pour estimer tous les paramètres du modèle. Ils mentionnent que ce problème est plus présent lorsque la taille du fichier de données d'analyse est grande.

Toujours dans le domaine de la santé publique, des chercheurs ont voulu utiliser des données de sources variées afin d'améliorer la qualité de modèles de survie. Ainsi, Horton *et al.* (2002) ont utilisé plusieurs sources de données afin de modéliser la participation d'un sujet à un rendez-vous de soin primaire après sa sortie de l'unité de désintoxication. Dans leur recherche, ils ont utilisé des données de sources variées afin d'intégrer une notion de délai entre la sortie de l'unité et le rendez-vous. Les auteurs mentionnent que toutes ces données ont été intégrées à même un modèle de survie multivarié. Selon eux, cette façon de procéder intégrait une méthodologie différente de ce qui avait été fait jusqu'à ce jour dans ce type de recherche. En effet, ils mentionnent

que lors de recherches antérieures, la participation à un rendez-vous de soin primaire était considérée comme réalisée si l'une ou l'autre des sources permettaient d'en arriver à cette conclusion. Avec cette nouvelle méthode, chaque source permettait, selon eux, de bonifier le modèle statistique par une meilleure estimation des paramètres.

La méthode appliquée par Horton *et al.* (2002) est très intéressante, puisqu'elle s'intéresse à l'intégration de toutes les données, et non une simple utilisation servant à préciser certains indicateurs. Dans le cas d'application qui nous intéresse ici, c'est exactement ce que nous souhaitons vérifier. Or, l'utilisation des sources de données externes est différente par le fait que dans l'étude de Horton *et al.* (2002), c'est pour la variable dépendante que ces données sont utilisées, alors que dans notre cas, c'est pour bonifier la partie explicative du modèle qu'elles sont utilisées.

Le cas d'application considéré dans la présente étude est relatif à la mesure de la performance de ventes en épargne dans un réseau de succursales du secteur bancaire. Or, la mesure de la performance n'est pas un sujet d'intérêt dédié uniquement au secteur financier. En effet, Higgins, Zeddies et Pearson (2011) se sont intéressés à la mesure de la performance de médecins individuels par l'utilisation de données provenant de multiples plans de santé. L'analyse des pratiques en santé de façon globale et comparative, selon ces auteurs, est difficile en raison des accès limités aux données, des règles touchant la protection des informations personnelles et de la complexité de comparaison de données provenant de sources diverses. En terme de protection des informations personnelles, les auteurs mentionnent que le problème se pose notamment lorsqu'il y a peu de patients pour un médecin. Afin de pallier ce problème de confidentialité, l'utilisation de données agrégées est une solution qui a été mise de l'avant par certaines initiatives locales de mesure de la performance en santé, selon Higgins, Zeddies et Pearson (2011). Ils ajoutent que, dans la majorité des cas, l'unité de mesure est le groupe de médecins et non le médecin individuel, mais que la mesure au niveau du médecin est très importante pour plusieurs raisons, telles que l'intérêt des

patients pour ce type d'information et la variabilité de la performance à ce niveau. Les auteurs ont utilisé une méthodologie qui se décline en deux étapes majeures :

1. déterminer les règles de mesure, en incluant les méthodes d'attribution des patients aux médecins; valider les résultats; établir les normes de présentation;
2. développer la solution technique de l'infrastructure de gestion des données qui serait utilisée pour l'agrégation des données provenant de différentes sources.

Dans leur recherche, Higgins, Zeddies et Pearson (2011) ont utilisé le nombre de visites d'un patient pour son attribution à un médecin. En ce qui a trait au choix de l'infrastructure de données, ils ont fait le choix d'un modèle réparti, étant donné les données confidentielles des patients qui ne pouvaient pas être déplacées et l'expertise locale devant être conservée concernant la manipulation et la préparation de données. En matière de mesures de la performance, ils se sont servis de calcul de moyennes, médianes et percentiles, ainsi que d'intervalles de confiance. Les auteurs suggèrent d'améliorer certains éléments par rapport à leurs travaux, comme l'utilisation d'un modèle plus rigoureux pour l'attribution des patients à un médecin (coût des soins accompagné du nombre de visites) et l'intégration d'autres sources de données à leurs données analysées, comme des données de santé mentale, de pharmacies et de laboratoires.

L'article de Higgins, Zeddies et Pearson (2011) est pertinent pour notre sujet d'intérêt pour plusieurs raisons :

- sujet de mesure de la performance dans un contexte de données agrégées;
- développement d'une méthodologie de consolidation des données provenant de différents secteurs;
- utilisation d'une méthodologie d'observations à un sujet d'intérêt;
- proposition d'améliorer les travaux par l'ajout de données de sources externes multiples.

1.3 Utilisation de données de sources multiples dans un contexte de données imprécises

Certains chercheurs se sont intéressés à l'utilisation de données de sources variées dans le cas de données imprécises. Ainsi, Baliga, Jain et Sharma (1997) ont discuté de ce type d'utilisation dans un cas d'apprentissage machine, en traitant de différents types de données imprécises. Cette recherche propose donc une méthodologie à utiliser dans ces types de cas. Toutefois, dans le cas d'application de notre projet de recherche, le problème de données imprécises n'est pas vécu, puisque les sources de données sont relativement propres et que le fait d'analyser des données agrégées diminue l'impact de certaines imprécisions au niveau du client.

1.4 Autres contextes d'utilisation de données de sources multiples

Nous l'avons vu, plusieurs cas d'utilisation de données de sources variées sont répertoriés dans le domaine de la santé publique. Cette pratique est toutefois remarquée dans d'autres domaines, comme en mesure du risque opérationnel, en hydrologie, ou en défaillances ou bris de systèmes. Dans le cas de défaillances de système, Reese *et al.* (2011) ont proposé un modèle intégrant l'information de différentes sources d'informations. Ces sources sont les suivantes :

- durées de vie des composantes individuelles;
- durées de vie du système ou des sous-systèmes;
- opinions d'experts sur des composantes spécifiques;

- opinions d'experts sur des groupes d'éléments.

Les chercheurs parlent moins dans ce cas des bénéfices ou des difficultés rencontrées lors de l'intégration de données de sources multiples, mais plutôt du modèle en tant que tel.

La prévision hydrologique est l'un des problèmes les plus importants des systèmes de ressources en eau, qui doit composer avec un traitement en temps réel, des alertes d'inondations et de sécheresse et l'irrigation planifiée (Azmi, Araghinejad et Kholghi, 2010). Dans ce contexte, Azmi, Araghinejad et Kholghi (2010) ont démontré que l'utilisation de données de sources multiples fusionnées peut améliorer de façon significative les prévisions, en comparaison à une utilisation de données de source unique. L'objectif principal de la fusion de données, définie par le processus de combiner ou d'intégrer des informations de multiples outils de mesures et/ou sources de données, est de fournir une solution plus précise relative à une mesure donnée, ou à faire des inférences supplémentaires par rapport à celles qui pourraient être obtenues grâce à l'utilisation des données de base (Dasarathy, B. V., 1997, cité par Azmi, Araghinejad et Kholghi, 2010). Dans le cadre de leur recherche, Azmi, Araghinejad et Kholghi (2010) ont comparé six méthodes de fusion de données. Ils ont démontré à l'aide de deux cas d'application que les prévisions effectuées à partir des données fusionnées surpassent celles des modèles individuels.

L'intérêt pour l'utilisation de sources externes de données en mesure du risque opérationnel semble assez marqué. En effet, « l'utilisation de données externes constitue une condition sine qua non dans l'implantation d'une méthode avancée de calcul de capital opérationnel, d'après les critères de Bâle II » (Dahen et Dionne, 2010). La mesure du risque opérationnel nécessite à la fois des connaissances techniques et commerciales d'outils quantitatifs, en plus de la compréhension des activités financières dans un sens très large (Bolancé *et al.*, 2013). Bolancé *et al.* (2013) mentionnent que les modèles de quantification du risque opérationnel ont peu de données à partir desquelles ils peuvent

être fondés. Selon Dahen et Dionne (2010), le recours à des données externes de pertes opérationnelles afin de compléter les données internes, en particulier dans le cas d'événements extrêmes qui ne sont pas présents dans les données internes, est essentiel. Toujours selon ces auteurs, la combinaison des données internes et externes d'une entreprise du secteur bancaire permet de réduire l'effet de surprise des événements extrêmes et de calculer le capital de risque opérationnel de façon adéquate. Les auteurs identifient quelques sources pertinentes dans le cas de l'évaluation du risque opérationnel :

- données publiques obtenues à partir des médias et de magazines;
- données provenant de courtiers d'assurances;
- données non publiques provenant de bases de données internes d'entreprises du secteur bancaire ayant accepté de partager ces informations.

Dahen et Dionne (2010) identifient certaines contraintes liées à chacune de ces sources de données, notamment des contraintes liées à la confidentialité des données et aux types de pertes considérées. Ces chercheurs identifient certains biais relatifs aux données externes, soient de sélection, de contrôle, de collection et de taille. Dans le cas de modèle d'évaluation du risque opérationnel, un biais statistique majeur provient du fait que les données externes sont tronquées au-dessus d'un seuil déterminé, alors que ce seuil peut être constant connu ou inconnu, ou stochastique (Baud, Frachot et Roncalli, 2002, 2007). Certains modèles d'évaluation du risque opérationnel donnent plus de poids aux données externes lorsque certaines données internes sont rares (Gustafsson et Nielsen, 2008).

Bien que l'évaluation du risque opérationnel ne soit pas le sujet de la présente étude, certains éléments peuvent être utilisés, notamment pour la compréhension de biais qui pourraient survenir en intégrant des données de sources variées.

Chapitre 2 : Cas d'application et données

2.1 Cas d'application

L'utilisation de sources multiples de données peut s'appliquer à de nombreux secteurs d'activité. Pour cette étude, une problématique liée au secteur bancaire est considérée. L'anonymat de cette entreprise devant être conservé, certaines informations seront brouillées, notamment en ce qui a trait aux informations sur l'entreprise, au nom des variables, aux résultats obtenus propres au domaine d'application et aux informations géographiques. Toutefois, il est souhaité que cette contrainte ne pénalise pas la compréhension du lecteur et son intérêt pour le sujet.

2.1.1 Entreprise et secteur d'activité

L'entreprise pour laquelle les données sont analysées est une entreprise majeure du secteur bancaire au Canada. L'entreprise possède plusieurs filiales et partenaires, et les clients peuvent entretenir une relation d'affaires avec une ou plusieurs de ces filiales ou partenaires. Dans cette étude, un réseau de distribution est considéré. Dans ce réseau, les services bancaires courants sont offerts, avec une offre de produits et services complets touchant l'épargne et le crédit. Ces services sont principalement destinés à une clientèle de masse, mais une clientèle aisée ou fortunée est également desservie par ce réseau. Ce réseau comprend quelques centaines de succursales (nombre confidentiel), qui sont situées sur un vaste territoire géographique au Québec et en Ontario.

L'entreprise possède une importante quantité de données portant sur ses activités et ces données sont emmagasinées dans plusieurs entrepôts de données ou bases de données, centralisés ou non dans l'entreprise.

2.1.2 Description du cas

Le cas d'application choisi concerne un secteur particulier des activités de l'entreprise, soit la distribution des produits d'épargne dans le réseau de distribution considéré. Le besoin est d'identifier les éléments permettant d'expliquer le mieux possible la performance des succursales en matière de ventes de produits d'épargne. Le volume d'affaires et la taille des succursales peuvent être très variés. Plusieurs produits sont considérés dans la grande famille des produits d'épargne et ces produits sont classés dans différentes familles de produits. La variété de produits offerts permet de répondre aux besoins d'une clientèle variée.

2.1.3 Approbation et confidentialité

Le secteur de l'entreprise concerné par cette analyse est en accord avec ce projet de recherche et avec l'utilisation de ses données, à la condition que le nom de l'entreprise ne figure pas dans le rapport, que toute information sensible ne soit pas divulguée et que l'interprétation des résultats respecte un certain niveau de confidentialité.

Toutes les données utilisées ne contiennent pas d'informations confidentielles sur des individus, ni même d'informations qui permettraient de les identifier. Cette affirmation est valide pour toutes les étapes du traitement de données.

2.2 Description des données

Dans ce projet de recherche, des données provenant de plusieurs sources sont utilisées. Ainsi, ces sources seront présentées dans les sections suivantes et les éléments relatifs à la période historique considérée, la fréquence de mise à jour et le niveau d'agrégation seront discutés.

2.2.1 Sources de données internes

Deux entrepôts de données de grande taille sont fréquemment utilisés dans l'entreprise pour alimenter les analyses, rapports de gestion, tableaux de bord et modèles statistiques. Ces entrepôts de données contiennent une importante quantité d'informations. Ces informations peuvent être relatives aux succursales (nombre d'employés, chiffre d'affaires, ventes et actifs reliés aux différents produits offerts, rentabilité, situation géographique, etc.), aux clients (informations sociodémographiques, comportement, appartenance, etc.) et aux produits et services offerts. Ces données servent d'importante base à l'analyse des activités de l'entreprise. Aucune donnée confidentielle sur les clients n'est présente dans ces entrepôts de données, puisque les identifiants sont créés de manière aléatoire et que toutes les informations permettant d'identifier des personnes sont retirées à la source par l'entreprise, avant leur intégration dans les entrepôts de données.

Afin de bonifier ces informations, d'autres sources de données internes à l'entreprise, mais non habituellement intégrées aux analyses, sont considérées. Dans le cas étudié, ces données sont de deux sources : l'outil de prestation de service utilisé dans toutes les succursales (un outil commun à toutes les succursales) et les informations de

participation de ces succursales relativement aux différentes activités de marketing direct et de sollicitation dans le temps.

Au sein de l'entreprise, un outil principal de prestation de service a été développé pour une utilisation dans toutes les succursales. L'utilisation de cet outil par les ressources travaillant au conseil et à la vente en succursales est obligatoire pour vendre certains produits de placement tels que des fonds de placement, mais facultatif pour la vente d'autres produits de placement. Or, il a été démontré par différentes analyses effectuées au sein de l'entreprise que l'utilisation de cet outil améliore les performances de ventes de produits d'épargne dans les succursales. Par contre, ces analyses ont été effectuées à partir de données collectées à même les ressources des succursales et, par hypothèse, ces données pourraient être différentes de ce qui est constaté en réalité par l'analyse des données de l'application. L'utilisation de cet outil permet de remettre au client une analyse de qualité d'une situation financière personnelle et des recommandations appropriées en matière d'épargne et de placement, ce qui rendrait le conseil au client crédible et adapté. Il est donc fortement souhaité que l'outil soit utilisé le plus souvent possible, mais de la bonne façon. Ainsi, certaines informations ont été extraites de l'application afin de construire des variables jugées potentiellement pertinentes par l'entreprise pour utilisation dans un modèle statistique. L'utilisation de ces nouvelles variables est possible étant donné les travaux récents effectués au sein de l'entreprise pour rendre ces informations disponibles, notamment pour effectuer des analyses statistiques.

Dans l'entreprise, des offensives tactiques concernant l'épargne et les placements sont développées pour les succursales. Ces offensives ciblent des clients en fonction de leur profil. Des offres spéciales, susceptibles d'intéresser ces clients, sont proposées. Toutefois, chaque succursale est en droit d'adhérer ou non à ces activités de sollicitation, lesquelles peuvent comprendre différents volets comme la télécommercialisation et des envois postaux. Ces données sont compilées à partir de rapports faisant état de l'adhésion

des succursales aux différentes activités. Cette compilation est faite dans le logiciel *Excel* de la suite *Microsoft Office*[1], où quelques variables d'intérêt sont créées, pour chacune des activités.

2.2.2 Sources de données externes

Les données de fréquentation du site web de l'entreprise sont accessibles par l'outil *Google Analytics*[2]. Différentes métriques sont disponibles et l'intégration de données de ce type aux données précédentes pour la construction de modèles statistiques est tentée. Ces données sont utilisées dans le but d'intégrer de l'information sur l'utilisation du site web de l'entreprise par les clients des différentes succursales. Ces données sont considérées comme externes dans cette recherche, puisque leur ajout dans des analyses statistiques constitue une nouveauté. De plus, l'angle sous lequel ces données sont utilisées, soit l'utilisation du web quant aux produits financiers sur le territoire des succursales, constitue une information externe aux informations de l'entreprise.

Finalement, des informations provenant de deux organismes externes, Statistique Canada et l'Institut national de santé publique du Québec (INSPQ; en collaboration avec le ministère de la Santé et des Services sociaux du Québec) sont ajoutées. Ces informations concernent l'emplacement géographique des succursales et le territoire qu'elles desservent. Il est important de préciser que l'Institut national de santé publique du Québec a étendu la portée des données qui seront utilisées à tout le Canada, ce qui permet de les utiliser dans ce cas d'application.

---

[1] http://office.microsoft.com/fr-ca/excel/
[2] http://www.google.com/analytics/

Les données de Statistique Canada utilisées proviennent du Fichier des attributs géographiques[3]. Dans ce fichier, des renseignements au niveau de l'îlot de diffusion, tels que des chiffres sur la population, les logements et la superficie des terres sont disponibles. Ce fichier comprend également des noms, classes et genres de codes géographiques. Finalement, les coordonnées géographiques associées aux aires de diffusion sont aussi incluses.

L'indice de défavorisation matérielle et sociale pour la santé publique a été développé pour le Québec à la fin des années 1990 et appliqué à l'ensemble du Canada en 2007 et 2008 par des chercheurs du ministère de la Santé et des Services sociaux et de l'Institut national de santé publique du Québec (Pampalon *et al.*, 2012). « L'objectif initial de l'indice était de pallier l'absence d'information socioéconomique dans les bases de données administratives du secteur de la santé et de décrire, à l'aide de ces bases de données, l'existence et l'ampleur des inégalités sociales de santé. » (Pampalon *et al.*, 2012). L'indice utilisé dans le cadre de ce cas d'application a été construit à partir des données du Recensement du Canada de 2006; il s'agit donc de la version 2006 de l'indice. Six indicateurs socioéconomiques ont été sélectionnés pour le développement de l'indice : « la proportion de personnes sans diplôme d'études secondaires; le ratio emploi/population; le revenu moyen personnel; la proportion de personnes vivant seules; la proportion de personnes veuves, séparées et divorcées; et la proportion de familles monoparentales » (Pampalon *et al.*, 2012). Ces indicateurs socioéconomiques ont été sélectionnés « pour leurs liens connus avec la santé et leurs affinités avec les deux dimensions de la défavorisation (matérielle et sociale) » (Pampalon *et al.*, 2012). Même si cet indice a été développé pour une utilisation relative à la santé publique, il appert que les indicateurs socioéconomiques qui le composent sont aussi pertinents pour une analyse de la clientèle d'une succursale sur un territoire donné, ce qui justifie son utilisation dans le cadre de notre cas d'application. En effet, des différences marquées

[3] Fichier des attributs géographiques,Recensement de 2011. Produit no 92-151-X au catalogue de Statistique Canada.

peuvent être observées pour certaines caractéristiques de la clientèle selon les territoires des succursales.

2.2.3 Historique, fréquence de mise à jour et niveau d'agrégation

Les données sont mises à jour de façon hebdomadaire, mensuelle ou trimestrielle, en fonction des différentes provenances. Elles sont disponibles avec un historique minimal de près de deux ans, mais dans certains cas, l'historique peut atteindre 10 ans. Afin d'avoir l'assurance que toutes les données seront disponibles aux différents moments considérés, la période d'analyse sélectionnée débute au début du quatrième trimestre de 2012 et se termine à la fin du second trimestre 2014; ces données sont compilées aux trimestres. De cette façon, un historique de sept trimestres (près de deux ans) est considéré dans les analyses.

Comme le cas d'application est relatif à la performance de succursales, toutes les données finales sont agrégées par trimestre et par succursale. Ainsi, puisque plusieurs centaines de succursales (nombre confidentiel) sont considérées pendant 7 trimestres, le jeu de données finales contient quelques milliers d'observations (nombre confidentiel).

Les sources de données ayant des niveaux d'agrégation variables, une partie importante de cette recherche sera relative aux méthodes utilisées pour fusionner les différentes sources de données. Dans ce document, une section complète est dédiée à cette partie, cruciale dans le cadre d'analyses statistiques de qualité et pour déterminer si l'intégration de données de sources multiples est une pratique avantageuse dans la construction de modèles statistiques.

Chapitre 3 : Méthodologie

3.1 Revue de la méthodologie de développement de modèles statistiques dans un contexte de mesures répétées

Lorsqu'une variable continue est à prédire, plusieurs méthodes statistiques et de *Data Mining* peuvent être considérées, telles que des arbres de décision, des réseaux de neurones, des modèles paramétriques comme la régression linéaire, des modèles semi-paramétriques ou d'autres techniques (Tufféry, 2010). La régression linéaire est sans doute la plus connue parmi ces techniques et est utilisée très fréquemment en modélisation statistique. Or, dans de nombreux cas de données réelles, certaines hypothèses de base nécessaires pour s'assurer de la validité des modèles ne sont pas vérifiées, dont l'hypothèse d'indépendance des observations (Larocque, Automne 2014). Le contexte de mesures répétées ou données longitudinales engendre ce type de situation lorsque plusieurs observations dans le temps sont prises sur des sujets d'intérêt, où les différentes observations pour un sujet sont possiblement corrélées (Larocque, Automne 2014). Dans ce contexte, le fait de tenir compte de cette dépendance est une condition nécessaire à la validité de l'analyse et cette condition se traduit par le fait de considérer une certaine structure de covariance dans le modèle (Larocque, Automne 2014).

Le modèle dans ce contexte peut s'écrire :

$$Y_{ij} = \beta_0 + \beta_{1ij}X_{1ij} + \cdots + \beta_{pij}X_{pij} + \in_{ij}$$

où :

- Y_{ij} est la valeur de la j^e observation du i^e sujet;

- X_{ij} est la valeur de la p^e variable pour la j^e observation du i^e sujet;

- $\in_{ij}$ est le terme d'erreur de la j^e observation du i^e sujet;

- β_0 est l'ordonnée à l'origine;

- β_p est le paramètre estimé de la p^e variable.

Pour modéliser la corrélation entre les observations pour un même sujet, une matrice de covariance pour ce sujet s'exprime de la façon suivante :

$$cov[Y_i|X_i] = \Sigma_i$$

3.1.1 Effets sur les paramètres estimés

Selon Larocque (Automne 2014), le fait d'ignorer une corrélation intrasujet dans le cas de mesures répétées se traduit par un impact sur l'estimation des écarts-types des paramètres estimés du modèle. Plus précisément, ignorer une corrélation intrasujet positive conduit à des tests qui rejettent l'hypothèse nulle trop souvent et à des intervalles de confiance trop courts.

3.1.2 Méthodes d'estimations des paramètres de la structure de covariance

Deux méthodes d'estimations des paramètres sont discutées par Larocque (Automne 2014), soient celles du maximum de vraisemblance (*Maximum Likelihood*, *ML*) et du maximum de vraisemblance restreint (*Restricted maximum likelihood, REML*). Selon l'auteur, lorsque le nombre de sujets est grand et le nombre de répétitions petit, il n'est

pas nécessaire de vérifier si les termes d'erreur d'un groupe obéissent à une loi multinormale. De plus, il mentionne que l'utilisation de la méthode *REML* est préférable pour l'estimation des paramètres de la structure de covariance par rapport à la méthode *ML*, afin d'éviter les biais sur les estimations, mais que ces biais sont pratiquement nuls lorsque la taille de l'échantillon est grande.

La procédure *PROC MIXED* de *SAS* permet d'utiliser l'une ou l'autre de ces deux méthodes d'estimation.[4]

3.1.3 Choix de la structure de covariance

Plusieurs structures de covariance sont disponibles avec la procédure *PROC MIXED* de *SAS*. Le Tableau 1 suivant résume ces différentes structures.

[4] http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_mixed_sect004.htm, consulté le 15 novembre 2014

Tableau 1 : Choix de structures de covariance offert par la procédure *PROC MIXED* de *SAS*

Structure	**Description**	**Paramètres**	**Éléments *(i, j)***
ANTE(1)	Ante-Dependance	$2t-1$	$\sigma_i\sigma_j\prod_{k=i}^{j-1}\rho_k$
AR(1)	Autoregressive(1)	2	$\sigma^2\rho^{\lvert i-j\rvert}$
ARH(1)	Heterogeneous AR(1)	$t+1$	$\sigma_i\sigma_j\rho^{\lvert i-j\rvert}$
ARMA(1,1)	ARMA(1,1)	3	$\sigma^2\left[\gamma\rho^{\lvert i-j\rvert-1}1(i\neq j)+1(i=j)\right]$
CS	Compound Symmetry	2	$\sigma_1+\sigma^2 1(i=j)$
CSH	Heterogeneous CS	$t+1$	$\sigma_i\sigma_j[\rho 1(i\neq j)+1(i=j)]$
FA(q)	Factor Analytic	$\frac{q}{2}(2t-q+1)+t$	$\sum_{k=1}^{\min(i,j,q)}\lambda_{ik}\lambda_{jk}+\sigma_i^2 1(i=j)$
FA0(q)	No Diagonal FA	$\frac{q}{2}(2t-q+1)$	$\sum_{k=1}^{\min(i,j,q)}\lambda_{ik}\lambda_{jk}$
FA1(q)	Equal Diagonal GA	$\frac{q}{2}(2t-q+1)+1$	$\sum_{k=1}^{\min(i,j,q)}\lambda_{ik}\lambda_{jk}+\sigma^2 1(i=j)$
HF	Huynh-Feldt	$t+1$	$\frac{\left(\sigma_i^2+\sigma_j^2\right)}{2}+\lambda 1(i\neq j)$
LIN(q)	General Linear	q	$\sum_{k=1}^{q}\theta_k A_{ij}$
TOEP	Toeplitz	t	$\sigma_{\lvert i-j+1\rvert}$
TOEP(q)	Banded Toeplitz	q	$\sigma_{\lvert i-j+1\rvert}1(\lvert i-j\rvert<q)$
TOEPH	Heterogeneous TOEP	$2t-1$	$\sigma_i\sigma_j\rho_{\lvert i-j\rvert}$
TOEPH(q)	Banded Hetero TOEP	$t+q-1$	$\sigma_i\sigma_j\rho_{\lvert i-j\rvert}1(\lvert i-j\rvert<q)$
UN	Unstructured	$t(t+1)/2$	σ_{ij}
UN(q)	Banded	$\frac{q}{2}(2t-q+1)$	$\sigma_{ij}1(\lvert i-j\rvert<q)$
UNR	Unstructured Corrs	$t(t+1)/2$	$\sigma_i\sigma_j\rho_{\max(i,j)\min(i,j)}$
UNR(q)	Banded Correlations	$\frac{q}{2}(2t-q+1)$	$\sigma_i\sigma_j\rho_{\max(i,j)\min(i,j)}$
UN@AR(1)	Direct Product AR(1)	$\frac{t_1(t_1+1)}{2}+1$	$\sigma_{i1j1}\rho^{\lvert i_2-j_2\rvert}$
UN@CS	Direct Product CS	$\frac{t_1(t_1+1)}{2}+1$	$\begin{cases}\sigma_{i1j1} & i_2=j_2\\ \sigma^2\sigma_{i1j1} & i_2=j_2\\ 0\leq\sigma^2\leq 1\end{cases}$
UN@UN	Direct Product UN	$\frac{t_1(t_1+1)}{2}+\frac{t_2(t_2+1)}{2}-1$	$\sigma_{1,i1j1}\sigma_{2,i2j2}$
VC	Variance Components	q	$\sigma_k^2 1(i=j)$ et i correspond au k^e effet

Toujours selon Larocque (Automne 2014), lorsque l'intérêt de la recherche concerne les effets des variables explicatives, la structure de covariance n'est pas intéressante pour analyse, mais doit tout de même être considérée étant donné la corrélation intrasuccursale pour le cas étudié. Ainsi, il mentionne que les critères *AIC* et *BIC* sont adéquats pour choisir cette structure et les plus petites valeurs obtenues dans ce cas sont associées aux meilleurs choix. De plus, l'auteur mentionne que lors du choix de la structure de covariance, les modèles doivent être comparés entre eux lorsqu'ils ont les mêmes effets fixes si la méthode *REML* a été utilisée, alors que toutes les comparaisons sont possibles dans le cas de l'utilisation de la méthode *ML*.

3.1.4 La sélection de modèles

Les méthodes de sélection de modèles classiques telles que *backward*, *forward* et *stepwise* ont été largement utilisées depuis plusieurs années. Ces méthodes sont facilement accessibles par le biais de logiciels statistiques tels que *SAS*. Or, à ces méthodes est associé un problème de sélection arbitraire des niveaux de signification qui permet de sélectionner ou d'exclure des variables en cours du processus de sélection, en plus des problèmes de tests multiples associés aux ajustements et réajustements du modèle (Bozdogan, 1987, Hosmer et Lemeshow, 1989, cités par Ngo et Brand, 2002). La question est d'autant plus complexe dans le cas de données répétées ou longitudinales, où la sélection du meilleur modèle ne signifie pas seulement choisir la meilleure structure de moyenne, mais aussi la structure de variance-covariance optimale (Wolfinger, 1996).

Récemment, Müller, Scealy et Welsh (2013) ont traité de la sélection de modèles pour les modèles linéaires mixtes. Ils mentionnent que ce type de modèles s'étant répandu dans les dernières années, le besoin d'avoir des méthodes et des outils de sélection a eu pour conséquence l'implantation d'un certain nombre de méthodes différentes de

sélection dans les logiciels (*R* ou *SAS*). Par contre, les auteurs constatent que quelques récentes méthodes n'ont pas encore été implantées dans les logiciels standards et qu'il n'y a pas de consensus dans la communauté statistique au niveau de la façon d'aborder la sélection de variables dans les modèles linéaires mixtes. Selon Müller, Scealy et Welsh (2013), comme les modèles linaires mixtes peuvent être vus comme étant des extensions de modèles de régression linaire, des extensions de méthodes développées pour la construction de modèles de régression linéaire peuvent être utilisées pour la sélection de variables pour les modèles linéaires mixtes. Ils mentionnent toutefois que cela ne signifie pas que la sélection de modèles pour les modèles mixtes linéaires peut être subsumée dans la sélection de modèles pour les modèles de régression linéaire. En effet, toujours selon Müller, Scealy et Welsh (2013), la dépendance entre les observations dans le cas d'un modèle linéaire mixte a un impact à la baisse sur la taille de l'échantillon, ce qui affecte les propriétés des procédures. Les méthodes de sélection de modèles étudiées par ces auteurs sont les suivantes : critères d'information (*AIC*, *BIC*), méthodes de réductions (*LASSO*), procédure de Fence et techniques bayésiennes. Müller, Scealy et Welsh (2013) ont émis plusieurs conclusions à partir de leurs recherches, dont celles-ci :

- la performance d'une méthode de sélection de modèles dépend de la façon dont elle est mesurée. Il est donc pertinent de considérer plusieurs mesures;

- la parcimonie est une considération importante lorsque le nombre de paramètres possibles est grand;

- lorsque le nombre de modèles à comparer est très grand, les ressources requises par des méthodes comme des critères d'information sont importantes. Il est donc judicieux de réduire le nombre de modèles à comparer. Les méthodes de réduction, de Fence et bayésiennes peuvent être meilleures dans le cas d'un grand nombre de modèles à comparer.

Dans un commentaire relatif aux recherches de Gurka (2006), Keselman *et al.* (2006) ont présenté plusieurs éléments pertinents à la sélection de modèles linéaires mixtes. Ils rapportent que l'auteur a déterminé à l'aide de simulations que les critères d'informations (*AIC*, *BIC*) sont très pertinents dans une sélection de modèles linéaires à mesures répétées. Keselman *et al.* (2006) mentionnent qu'ils ont choisi, afin d'apporter quelques précisions aux travaux de l'auteur, de trouver tout d'abord la structure de covariance appropriée avec les critères d'information *AIC* et *BIC* et d'utiliser ensuite cette structure pour effectuer la sélection des effets fixes. Ils mentionnent avoir choisi cette approche puisqu'en utilisant la structure de covariance appropriée, une meilleure précision dans les tests des mesures répétées est constatée. Finalement, Keselman *et al.* (2006) mentionnent que, selon leur expérience sur des données réelles (non simulées), la structure de covariance non structurée est la plus pertinente et que la recherche selon les critères d'information ne vaut souvent pas les efforts investis.

Ngo et Brand (2002) recommandent l'utilisation du critère d'information Akaike (*AIC*) pour la sélection de modèles linéaires pour données répétées ou longitudinales, afin d'éviter de devoir spécifier un seuil de signification arbitraire dans une méthode de sélection automatisée telle que la méthode *stepwise*. Les auteurs proposent deux méthodes permettant de sélectionner le modèle présentant la plus faible valeur de *AIC* :

1. identifier toutes les combinaisons possibles de variables et toutes les structures de variance-covariance applicables à la question d'intérêt. Le nombre de modèles considérés inclut toutes les combinaisons possibles de variables avec les structures de variance-covariance. Pour chaque modèle, l'*AIC* est calculé et le modèle associé au plus petit *AIC* est sélectionné;

2. utiliser la structure la plus complexe considérée et sélectionner la meilleure structure de variance-covariance pour ce modèle. On peut calculer le AIC_R selon la méthode du maximum de vraisemblance restreint (*REML*) :

$$AIC_R = -2 * (Ressemblance\ restreinte) + 2 * (nombre\ de\ paramètres\ de\ covariance)$$

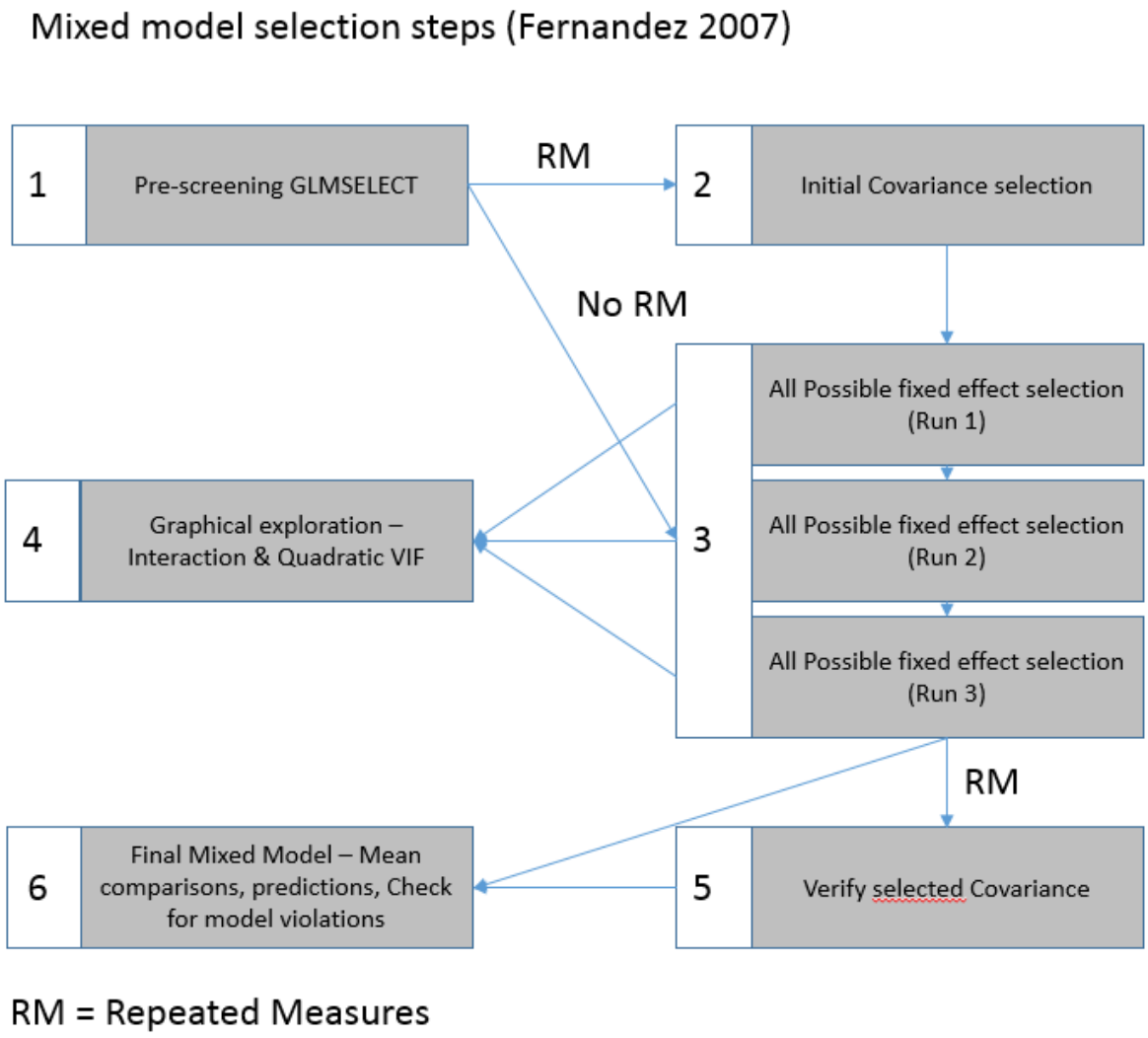
La structure de variance-covariance présentant la plus petite valeur de AIC_R est sélectionnée. Utiliser le *AIC* afin de sélectionner le meilleur modèle en utilisant cette structure de variance-covariance.

Fernandez (2007) a proposé une macro à utiliser dans le logiciel *SAS* afin de procéder à une sélection de modèles avec effets fixes en présence de mesures répétées en utilisant *SAS PROC MIXED*. Plusieurs options sont proposées par l'auteur :

- sélection de la meilleure structure de covariance;
- exploration graphique et détection des effets fixes significatifs;
- diagnostiquer la multicolinéarité.

Cette macro a l'avantage de rendre disponible une solution complète de sélection de variables automatisée, qui tient compte de plusieurs étapes clés, selon l'auteur. Les étapes clés, selon Fernandez (2007), sont présentées à la Figure 1.

Figure 1 : Étapes de sélection de modèles et leur séquence recommandée (Fernandez, 2007)



Dans son article, Fernandez (2007) décrit toutes les étapes considérées dans la macro SAS qu'il a développée (*ALLMIXED2*). Cette description détaillée est très pertinente pour le cas d'application considéré dans cette recherche, que la macro soit utilisée ou non. Voici donc un résumé des différentes étapes considérées et des choix qu'il a faits lors de la construction de la macro *ALLMIXED2*. Toutes ces étapes sont proposées par Fernandez (2007).

Étape 1 : Présélection de variables

Lorsque le nombre de variables est important, l'auteur mentionne qu'une présélection de variables est requise, puisque le temps et les ressources de traitement lors de l'évaluation des modèles possibles seraient trop importants. Cette présélection sert à éliminer les termes peu contributeurs. Lors de cette étape, la structure de covariance des mesures répétées est ignorée et les effets aléatoires peuvent être considérés comme fixes. L'auteur recommande la méthode *LASSO* (Tibshirani, 1996, cité par Fernandez, 2007), qui a été implémentée dans la procédure *GLMSELECT* de *SAS* (Cohen 2006, cité par Fernandez, 2007). Les options relatives à la méthode *LASSO* dans la procédure *GLMSELECT* de *SAS* sont *CHOOSE=NONE* et le critère de sélection *SBC*. Il recommande l'utilisation des graphiques « *FIT CRITERIA* » et « *COEFFICIENT EVALUATION* » disponible dans *SAS ODS GRAPHICS* afin d'identifier le potentiel de sous-ensembles possibles, afin d'éliminer les covariables présentant peu de potentiel et afin de sélectionner le nombre de variables significatives requis (l'auteur parle d'un maximum de 10).

Étape 2 : Sélection de la structure de covariance des mesures répétées

La sélection de la meilleure structure de covariance est effectuée en comparant les statistiques *AICC* obtenues à l'aide de la méthode *PROC MIXED REML* de *SAS*, où :

$$AICC = -2\log L + \left[2(p+k+1)\left(n/(n-p-2)\right)\right];$$

p = nombre d'effets fixes;

k = nombre d'effets aléatoires;

n = nombre de sujets.

La plus faible valeur de la statistique *AICC* détermine le meilleur choix de structure de covariance.

Étape 3 : Sélection de modèles en considérant tous les modèles possibles

L'auteur suggère de comparer les statistiques *AICC* et *MDL* (où $MDL = 1/2\{-2logL + [\log(n)\,(p + k + 1)]\}$) , à cette étape, pour toutes les combinaisons d'effets fixes. Il explique que l'utilisation de la statistique *AICC* favorise des modèles parcimonieux, alors que la statistique *MDL* favorise des modèles avec un plus grand nombre d'effets fixes, surtout dans le cas de petits échantillons. Toutefois, lorsque le nombre d'observations est grand, la statistique *MDL* semble favoriser des modèles plus parcimonieux. On cherche les plus faibles valeurs de ces deux statistiques. L'auteur mentionne que les meilleurs candidats de modèles sont ceux pour lesquels la statistique AICC est à une différence plus faible ou égale à deux par rapport à la plus petite valeur de cette statistique obtenue en comparant tous les modèles ($\Delta AICC = AICC_i - AICC_{min}$). De la même façon, les meilleurs candidats de modèles sont ceux pour lesquels la statistique *MDL* est à une différence plus faible ou égale à un par rapport à la plus petite valeur de cette statistique obtenue en comparant tous les modèles ($\Delta MDL = MDL_i - MDL_{min}$).

Étape 4 : Exploration graphique pour détection de multicolinéarité et spécification de l'erreur du modèle.

Selon l'auteur, la présence de multicolinéarité sévère entre les variables explicatives dans l'analyse de modèles mixtes peut conduire à des estimations de paramètres instables avec des erreurs standards gonflées. Cette présence peut affecter la sélection des effets fixes. Ainsi, il est préférable de sélectionner des variables qui ne sont pas associées à de la multicolinéarité. L'auteur suggère une exploration graphique présentant le facteur d'inflation de la variance (*VIF*) à l'aide de box-plot. Il suggère de se soucier des valeurs de *VIF* supérieures à 10.

Toujours selon l'auteur, le succès de l'étape de sélection de modèles peut être compromis par l'omission de termes significatifs d'ordre supérieur. Afin de valider si un terme significatif d'ordre quadratique doit être ajouté au modèle, l'auteur recommande d'ajuster le modèle en incluant le terme, d'examiner si le seuil observé de la statistique

de *type III* est significatif et de conserver les valeurs prédites $\hat{Y}$ pour ce modèle complet. Par la suite, le modèle excluant à la fois les termes linéaire et quadratique pour un prédicteur est ajusté et les valeurs prédites $\hat{Y}$ de ce modèle réduit sont conservées. La représentation graphique de la valeur $\hat{Y}_{complet} - \hat{Y}_{reduit}$ en fonction des valeurs du prédicteur révélera la nature et la force de l'effet quadratique. La même procédure peut être répétée pour les interactions entre deux prédicteurs, en utilisant un graphique en trois dimensions pour comprendre la nature et la force de l'effet d'interaction.

Étape 5 : Sélection finale de la structure de covariance

L'auteur recommande de comparer des modèles avec plusieurs structures de covariance en utilisant la statistique $\Delta AICC = AICC_i - AICC_{min}$ obtenue de la méthode *PROC MIXED REML* de *SAS*.

Étape 6 : Analyse du modèle mixte sélectionné

L'auteur suggère finalement de procéder à l'exploration de données à l'aide de box-plot, d'analyser le modèle mixte, d'effectuer des comparaisons à l'aide de *LSMEAN*, de faire des prédictions, de faire la vérification au niveau de la normalité des résidus conditionnels studentisés et d'effectuer des diagnostics d'influence.

3.2 Méthodologie utilisée

3.2.1 Méthodes statistiques sélectionnées

Les modèles qui seront considérés pour vérifier l'hypothèse de recherche sont des modèles linéaires mixtes avec mesures répétées (données longitudinales). Effectivement, comme les données sont prises à sept reprises sur une période de temps pour les succursales, il importe de tenir compte d'une situation de dépendance entre certaines observations. Ainsi, les modèles tiendront compte d'une corrélation à l'intérieur de chaque groupe défini par une succursale. Aucune corrélation intra secteur ou intra sous-secteur n'est toutefois considérée, puisque l'hypothèse que ces éléments de structure organisationnelle n'ont aucun impact sur la performance des succursales est posée.

4.3.2 Méthodologie de sélection de modèles

Étant donné le contexte de sélection de modèles dans un contexte de mesures répétées, une méthodologie appropriée est développée. Cette méthodologie est légèrement inspirée de Fernandez (2007), Keselman *et al.* (2006) et Ngo et Brand (2002), mais présente plusieurs nouveautés étant donné le très grand nombre de variables considérées. L'efficacité de cette méthode sera discutée au chapitre suivant.

Étape 1 : Présélection de variables

Effectuer une présélection de variables en ne tenant pas compte du contexte de mesures répétées. Cette présélection est faite en *SAS* à l'aide de la procédure *GLMSELECT*. La méthode *STEPWISE* traditionnelle basée sur la statistique F est utilisée, avec de larges seuils de 0,4 en entrée et en sortie. En effet, des critères peu restrictifs sont préférés de

ce point de vue afin de réduire le nombre de variables à moins de 50. Le critère *AIC* est choisi pour la sélection finale au profit d'autres critères étant donné que le but de cette étape est de conserver un plus grand nombre de variables potentielles et que les critères *SBC* et *BIC* sont plus restrictifs. Les méthodes de sélection *LASSO* et *LAR*, parfois recommandées dans la littérature, sont aussi jugées trop restrictives pour cette étape de présélection.

La sélection de variables est effectuée en tenant compte des variables catégorielles, telles que le secteur, le sous-secteur et le trimestre. Pour ce faire, l'énoncé CLASS est utilisé et permet de considérer un effet fixe pour chaque catégorie de ces variables.

Étape 2 : Choix de la structure de covariance

Le choix initial de la structure de covariance est fait en fonction d'un modèle avec tous les effets fixes sélectionnés à l'étape 1. Les critères de sélection *AIC* et *BIC* sont utilisés afin de déterminer la meilleure structure. La structure associée aux plus faibles valeurs de ces critères est la plus appropriée à cette étape. Le choix de la structure de covariance est fait à l'aide de la procédure *PROC MIXED*. Étant donné la comparaison de modèles comprenant le même nombre d'effets fixes, la méthode *Restricted Maximum Likelihood* (*REML*) est utilisée.

Étape 3 : Sélection des effets fixes

En tenant compte de la structure de covariance sélectionnée à l'étape 2, une recherche des meilleurs modèles par nombre d'effets fixes considérés est réalisée. Les effets fixes potentiels considérés à cette étape ne comprennent pas d'interactions ou d'effets d'ordres supérieurs, mais bien des effets fixes identifiés à l'étape 1 de présélection.

Pour cette troisième étape, le contexte de mesures répétées est tenu en compte, c'est pourquoi les méthodes de sélection disponibles dans des procédures telles que *REG* ou *GLMSELECT* ne peuvent être utilisées. Pour cette raison, j'ai développé un algorithme de sélection permettant d'utiliser la procédure *MIXED* (méthode *ML*). Cet algorithme de sélection permet de tenir compte de la structure de covariance entre les mesures répétées lors de la comparaison d'un très grand nombre de modèles. Elle se veut être un juste milieu entre une méthode de sélection par la comparaison de tous les modèles possibles, impossible en tenant compte des mesures répétées avec les ressources informatiques disponibles, et une méthode de type *STEPWISE*. Des contraintes sont utilisées afin de limiter le nombre de modèles à comparer. Voici les différentes étapes proposées dans cette méthode :

1. trouver le meilleur modèle avec un effet fixe en conservant les valeurs des critères de sélection *AIC* et *BIC* pour tous les modèles à un effet dans un fichier global. Ces modèles sont ajustés un à un de façon automatisée. Ne conserver que le modèle associé à la plus faible valeur de *AIC* ou *BIC*. Puisque le nombre de variables est le même pour tous les modèles considérés, les deux critères recommanderont toujours le même modèle;

2. trouver le meilleur modèle avec deux effets fixes de façon similaire à l'étape 1. Toutes les combinaisons possibles différentes de deux effets fixes sont ajustées de façon automatisée. Ne conserver que le modèle à deux effets fixes associé à la plus faible valeur de *AIC* ou *BIC*;

3. si un même effet fixe a été sélectionné dans les meilleurs modèles à un et deux effets fixes (à l'issue des deux étapes précédentes), fixer cet effet fixe afin de restreindre le nombre de possibilités de modèles à trois effets fixes. Si aucun effet fixe n'a été sélectionné autant à l'étape 1 qu'à l'étape 2,

considérer toutes les possibilités différentes de modèles à trois effets fixes. Ajuster toutes les possibilités de modèles à trois effets fixes, dont l'un fixé ou non, et ne conserver que le modèle associé à la plus faible valeur de *AIC* ou *BIC*;

4. répéter l'étape 3 afin d'identifier les meilleurs modèles à *i* effets fixes, où *i* est égal ou supérieur à 4, mais ne dépasse pas le nombre total d'effets présélectionnés moins 1. Si, à l'issue de l'étape précédente, aucun effet ne peut être fixé selon les conditions mentionnées (très rare, mais possible), fixer un effet sur la base du fait qu'il avait déjà été sélectionné dans un modèle à moins d'effets, mais pas lors de deux processus conjoints;

5. comparer les meilleurs modèles associés à tous les nombres d'effets fixes et conserver celui présentant la plus faible valeur de *AIC* et celui présentant la plus faible valeur de *BIC*. Une fois que ces critères semblent avoir atteint des minimums et que les meilleurs modèles de quelques processus suivants ne semblent pas être plus performants, l'étape de sélection peut se terminer.

La macro *SAS* développée pour effectuer la sélection de modèles est présentée en Annexe 1. La méthode *ML* est utilisée étant donné le nombre d'effets fixes variables dans les modèles comparés.

Étape 4 : Validation du meilleur modèle

Quelques vérifications sont effectuées afin de s'assurer que le modèle choisi est adéquat pour une bonne interprétation. Tout d'abord, une exploration graphique des résidus permet de valider s'ils sont de moyenne nulle. Ensuite, des graphiques présentant la distribution des résidus par rapport aux différentes variables explicatives sont visualisés afin de détecter la présence d'hétéroscédasticité. L'option SPEC est utilisée à titre

indicatif afin de voir si, de façon globale, l'hypothèse nulle d'homoscédasticité est rejetée. En cas de présence d'hétéroscédasticité, des transformations de variables sont tentées pour enrayer le problème. En ce qui a trait à la validation de la normalité de la distribution des erreurs, le *QQ-plot* est analysé. Finalement, la détection de problème de multicolinéarité est effectuée par l'utilisation de l'option *COLLINOINT* de la procédure *REG*. Plus spécifiquement, les grandes valeurs associées aux mesures « *Condition Index* » et « *Variance Proportion* » sont suspectes. Dans le cas présent, des valeurs supérieures à 10 seront inquiétantes. En cas de multicolinéarité, des variables problématiques pourront être retirées à l'étape de leur entrée dans le modèle au profit d'autres variables. Certaines étapes du processus de sélection de modèles pourraient donc être reprises en considérant les nouvelles informations sur la relation entre les variables explicatives.

Comme les modèles sont construits dans un but d'expliquer la performance des succursales, les modèles choisis au terme de l'étape 4 sont censés être de bons modèles. Or, il est possible d'aller plus loin et de considérer l'ajout de termes d'ordres supérieurs et d'interactions entre plusieurs variables. En intégrant de tels effets fixes, les modèles peuvent toutefois se complexifier rapidement et rendre l'interprétation des paramètres des effets fixes compliquée, voire impossible, surtout dans un contexte de modèles explicatifs. Une méthode d'intégration de tels effets dans un but d'améliorer la performance des modèles est présentée à l'étape 5, mais cette étape n'est qu'un préambule pour aller plus loin. Les modèles avec interactions et termes d'ordre supérieurs ne seront pas utilisés dans l'analyse des résultats. De tels modèles seraient toutefois utiles dans un but de prédiction de la performance des succursales.

Étape 5 : Ajout d'interactions et de termes d'ordre supérieurs (pour aller plus loin)

Rechercher le meilleur modèle avec un terme d'ordre supérieur. Pour ce faire, considérer toutes les possibilités d'interactions entre deux variables sélectionnées aux termes des étapes 3 et 4 et toutes les possibilités de ces variables au carré. Ajuster tous les modèles

avec l'ajout d'un effet d'ordre supérieur et conserver celui présentant la plus faible valeur associée au critère *BIC*. Fixer ce nouvel effet fixe et recommencer pour identifier le meilleur modèle avec deux effets fixes d'ordre supérieur. Poursuivre jusqu'à ce que ce processus ne permette plus de diminuer la valeur du critère *BIC* pendant quelques étapes. Le modèle associé à la plus petite valeur du critère *BIC* est le modèle choisi. Finalement, la structure de covariance est révisée afin de valider si elle est toujours celle qui convient le mieux.

Chapitre 4 : Préparation des données

4.1 Intégration des données de sources multiples

Au chapitre 1, nous avons vu que l'intégration des données consiste en l'élaboration d'un processus qui permet de consolider, au sein d'une même infrastructure de données, des informations provenant de sources multiples. Même si un processus d'intégration est habituellement vu comme étant récurrent et structuré, les étapes de manipulations de données provenant de différentes sources concernant notre cas d'application peuvent être vues comme un exercice d'intégration non récurrent afin de répondre aux besoins d'analyse déterminés au préalable, soit d'expliquer la performance d'une succursale en regard de plusieurs éléments mesurés. Par cet exercice, il sera possible de mettre en évidence les efforts qui lui sont associés. De plus, certaines préoccupations liées à un processus d'intégration de données doivent être interrogées afin de mener à bien cette activité déterminante pour le développement de modèles statistiques adéquats. Ainsi, les paragraphes suivants font état de l'élaboration du processus pour le cas d'application considéré en considérant certains éléments importants.

Tout d'abord, nous avons vu dans une section précédente que des sources de données ont été sélectionnées pour répondre à une question liée à notre cas d'application. En effet, ces choix résultent d'une réflexion concernant les différents éléments qui pourraient affecter la performance d'une succursale. La connaissance du secteur d'activité de l'entreprise, des discussions avec différents professionnels et la consultation de résultats publiés en silo au sein de l'entreprise ont permis de définir un sous-ensemble de sources intéressantes pour ce cas d'application. De plus, un intérêt

marqué pour certains types de données, comme celles provenant de la mesure du web, a influencé le choix des sources.

Ensuite, l'analyse de la qualité des données est un élément important, puisque la performance des modèles développés peut être fortement affectée dans le cas d'une qualité non adéquate. Comme les sources sont multiples, il importe d'aborder la qualité des données de façon isolée en fonction de chaque source.

Les données provenant des entrepôts de données internes de l'entreprise proviennent pour la plupart de systèmes opérationnels. Des règles d'affaires et des modèles sont appliqués aux données brutes afin de construire des variables et indicateurs pertinents et cohérents. Ainsi, un certain nettoyage est effectué à partir de la source, ce qui garantit un niveau de qualité satisfaisant. Du point de vue du client, des données sont manquantes pour plusieurs indicateurs, surtout ceux de nature non financière. Toutefois, comme ces données sont agrégées au niveau de la succursale pour les analyses, les valeurs manquantes observées au niveau du client ont été conservées et permettent la construction de variables lors de l'agrégation. Aucune imputation n'est effectuée pour les données des entrepôts afin de conserver l'information de donnée manquante qui pourrait être importante dans les modèles.

Les données de Statistique Canada et de défavorisation pour la santé publique ne semblent pas comporter d'irrégularités. En effet, tous les secteurs des succursales évaluées trouvent une correspondance dans les fichiers de données de ces deux organismes et aucune donnée n'est manquante pour les variables choisies.

La qualité des données relatives à la prestation de service semble bonne. En effet, les indicateurs proviennent d'un traitement effectué à partir du système source, où des règles d'affaires sont appliquées. Une absence de données de prestation de service pour un

client en particulier signifie simplement que l'outil n'a pas été utilisé pendant la période pour ce client.

Puisque les données relatives à la participation des succursales à des activités de sollicitation proviennent de fichiers individuels pour chaque offensive, la qualité des données pourrait être moins bonne que dans le cas des sources de données considérées dans les paragraphes précédents. En effet, les données sources proviennent de fichiers de formats variés, où les données sont entrées à la main et par des employés différents. De plus, ces informations ne sont pas conservées dans un but d'analyse, mais plutôt afin de garder une trace des activités vécues dans le temps. Toutefois, il n'y a pas de façon de vérifier la validité des données. Il faut donc se fier aux informations compilées. Il n'y a pas beaucoup de données manquantes pour les données de cette nature.

Le principal défi sur le plan de la qualité des données est assurément associé aux données extraites de l'outil *Google Analytics*. Effectivement, l'outil est utilisé dans un but de suivi global et d'analyse à haut niveau. Le fait d'utiliser ces données, agrégées et sur lesquelles des règles d'affaires ont été appliquées, dans un but d'apporter plus d'informations dans des modèles statistiques comporte certains défis en lien, notamment, avec le jumelage des observations. De plus, certaines informations sont brouillées par l'entreprise *Google* afin de s'assurer du respect de la confidentialité des données de certains de leurs utilisateurs, notamment de clients d'autres filiales de *Google*. Ainsi, plusieurs observations correspondent à une ville inconnue et ne peuvent donc pas être associées à des succursales. De plus, le jumelage des villes aux succursales comporte de nombreuses contraintes qui seront discutées plus loin, dans la section liée à l'intégration de cette source de données.

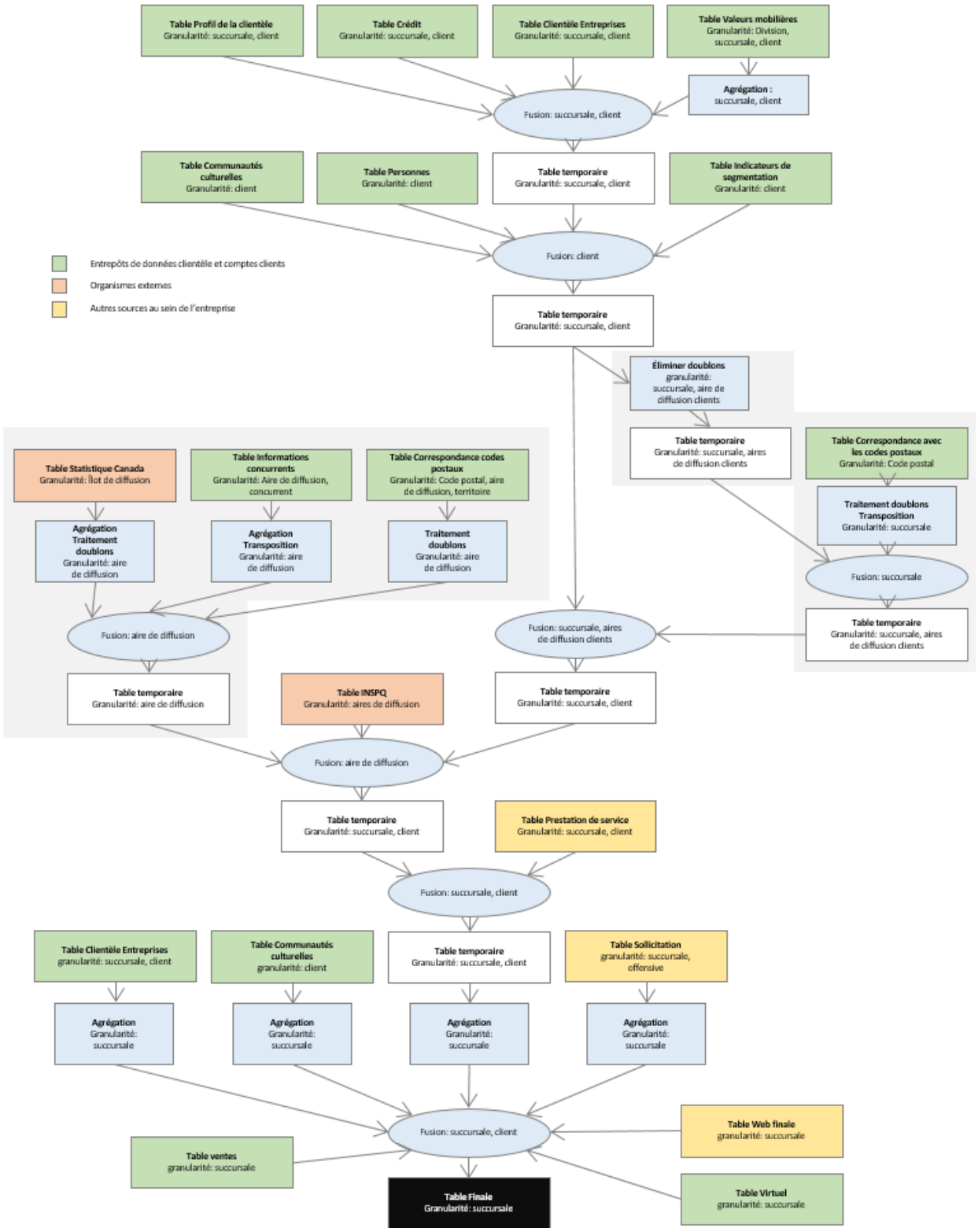
La Figure 2 (une vue scindée est présentée en annexe 2) résume, à l'aide d'un diagramme, les différentes étapes développées dans le cadre de ce projet de recherche permettant de lier les données provenant des multiples sources entre elles. Il est

important de mentionner que les étapes de manipulation de données ont été résumées et simplifiées afin de permettre une vue d'ensemble des traitements. Les détails relatifs aux tables de données concernées sont présentés dans les sous-sections suivantes, le diagramme ayant pour objectif premier de constater, à haut niveau, les manipulations effectuées. Le secteur de l'entreprise concerné par ce projet de recherche n'a pas établi de structure d'intégration de ce type à ce jour, puisque plusieurs tables de données des entrepôts internes ne sont pas utilisées à des fins d'analyse de la performance en épargne. En effet, les travaux intégrant des variables de nature sociodémographique ne sont pas fréquents dans ce secteur. En raison de la nature confidentielle des données (nom des variables et des tables de données), les programmes *SAS*[5] développés dans le cadre du processus d'intégration des différentes sources de données ne peuvent être présentés.

Les étapes d'intégration des données de sources multiples ont principalement été effectuées à l'aide du logiciel *SAS*, version 9.4. Par contre, pour certaines étapes, le logiciel *Excel* de la suite Microsoft Office 2013 a été nécessaire.

[5] http://www.sas.com/fr_ca/home.html

Figure 2 : Résumé du traitement proposé à l'entreprise afin de lier les différentes sources de données (vue scindée en annexe 2)



4.1.1 Entrepôts de données clientèle et comptes clients

Les entrepôts de données clientèles et comptes clients sont des cas d'intégration des données, tel que défini dans le chapitre 1 par les différents auteurs cités. En effet, les données proviennent de sources multiples et sont intégrées à l'aide de processus structurés. Ces données sont transformées afin de répondre aux normes de l'entrepôt de données, particulièrement en ce qui a trait à leur format. Un horaire de mise à jour est également respecté. Ainsi, l'utilisation isolée de ces données constitue déjà un exercice d'utilisation de données de sources multiples, mais un des objectifs de cette recherche est de vérifier si le fait d'aller plus loin dans cette direction constitue une amélioration significative. Conséquemment, comme il s'agit de données internes structurées et entreposées, ces données seront considérées comme provenant d'une source unique.

L'entrepôt de données sur la clientèle contient les informations de base qui seront utilisées dans les analyses statistiques. Dans l'entreprise, ces informations sont déjà utilisées dans certains secteurs pour la construction de modèles statistiques. Ainsi, ces données peuvent servir en ciblage de clients pour des activités de marketing direct, pour la construction d'indicateurs sur le profil des clients ou pour créer des segments qui permettent de mieux comprendre le comportement des clients. Toutefois, pour l'analyse de la performance des différentes succursales en lien avec les ventes en épargne et placement, ces données sont très peu utilisées et il n'existe pas de modèle explicatif ou prédictif en fonction des caractéristiques des clients. Pour le moment, le suivi de la performance des succursales se fait donc seulement sur la base de variables simples comme le volume d'affaires, sans tenir compte de certains contextes particuliers, tant au niveau géographique qu'au niveau de la clientèle desservie.

La sélection des différentes informations à être utilisées pour la construction des modèles s'est effectuée de façon rigoureuse, en considérant plusieurs éléments pouvant être déterminants dans la différentiation par le contexte d'affaires des différentes succursales. Ainsi, les données ont été sélectionnées dans 11 tables de données. Ces données ont été obtenues par des requêtes *SQL* sur des vues via un entrepôt de format *DB2*. Les requêtes *SQL* sont lancées depuis une connexion *SAS* en *RSUBMIT*, à l'aide de la procédure *SQL*. Ces requêtes ne sont pas présentées étant donné le niveau de confidentialité exigé. Le Tableau 2 donne la description des tables de données sur lesquelles l'extraction est bâtie.

Tableau 2 : Description des tables de données des entrepôts de données clientèle et comptes clients

	Nom	**Description**	**Granularité**
Table1	Profil de la clientèle	Collection d'informations en vue de présenter des éléments de profilage notamment sur les actifs, les différents produits détenus ainsi que les indicateurs informatifs et les données transactionnelles.	Succursale, client
Table2	Personnes	Contient une série d'indicateurs permettant de faire un profil sociodémographique et de déterminer la relation d'une personne cliente avec l'entreprise.	Client
Table3	Indicateurs de segmentation	Regroupe les résultats de plusieurs modèles de segmentation et modèles prédictifs.	Client
Table4	Correspondance avec les codes postaux	Contient, par aire de diffusion et par territoire de succursales, les informations sur les codes postaux et leurs coordonnées géographiques (latitude et longitude).	Code postal, aire de diffusion, territoire

Table5	Informations concurrents	Information des concurrents par aire de diffusion. Comprends l'adresse (y compris latitude et longitude) et le nom du concurrent.	Aire de diffusion, concurrent
Table6	Communautés culturelles	Information indiquant si le client fait partie d'une communauté culturelle et laquelle.	Client
Table7	Crédit	Informations des clients relativement aux produits de crédit.	Succursale, client
Table8	Valeurs mobilières	Contient de l'information sur la clientèle et les comptes de valeurs mobilières.	Division, succursale, client
Table9	Clientèle Entreprises	Regroupe des données au sujet des clients "Entreprise".	Succursale, client
Table10	Compte client	Contient des informations spécifiques et disponibles au niveau du compte du client.	Client, compte
Table11	Sites physiques	Permet d'identifier tous les sites physiques utilisés, ou qui ont été utilisés, dans l'entreprise. Les informations que l'on y retrouve permettent d'identifier le type de point de services, son statut, le nom et l'adresse du point de service et ses informations de géolocalisation.	Succursale, point de service

4.1.2 Données de Statistique Canada

Les données de l'étude proviennent toutes du *Fichier des attributs géographiques*, pour l'année de recensement 2011. Ces données sont utilisées afin de tenir compte du contexte de l'emplacement géographique de la succursale et des clients. Ainsi, ces informations sont utilisées à deux niveaux différents : le territoire de la succursale et l'emplacement

de la résidence du client. Les variables jugées pertinentes pour utilisation dans le cas d'application sont les suivantes :

- population de l'îlot de diffusion selon le Recensement de 2011 (1);
- total des logements privés de l'îlot de diffusion selon le Recensement de 2011 (2);
- logements privés occupés par les résidents habituels de l'ilot de diffusion selon le Recensement de 2011 (3);
- aire de diffusion (4);
- latitude et longitude du point représentatif de l'aire de diffusion, en degrés et décimales (5);
- province ou territoire (6);
- genre de la subdivision de recensement (7);
- genre du centre de population et région rurale (8);
- classification des centres de population et régions rurales (9).

Ces variables visent à ajouter de l'information par rapport, notamment, à la densité de population dans le secteur desservi par la succursale ou dans l'aire de diffusion du client, aux coordonnées géographiques des différents lieux et à la situation rurale ou urbaine d'un lieu.

Certaines manipulations sur les données sont effectuées au départ afin d'être en mesure de les combiner aux données par succursale ou par client. Tout d'abord, les données, en format *Excel*, sont importées dans *SAS* à l'aide d'une étape *DATA*. Seules les informations décrites plus haut sont conservées, pour les provinces de Québec et de l'Ontario. Puisque les données sont agrégées à un niveau plus fin que l'aire de diffusion, il importe d'agréger les informations pertinentes sous le format d'une ligne par aire de

diffusion. Ainsi, les données sur la population et sur les logements (1, 2, 3) sont agrégées par aire de diffusion, alors que les informations relatives aux plus grands nombres d'habitants sont conservées en ce qui a trait aux centres de population et régions rurales (8, 9). Cette information présentée dans un nouveau format peut ensuite être utilisée par succursale et par client.

L'ajout de ces informations par succursale permet de constater la taille de la population dans les aires de diffusions desservies par les succursales et les logements associés à ce territoire. C'est par les aires de diffusions que s'effectue l'association des informations pertinentes à chaque succursale. En effet, à chaque succursale est associé un territoire, représenté par une ou plusieurs aires de diffusions. Ainsi, la fusion des données géographiques se fait par l'aire de diffusion. Ensuite, la somme de toutes les observations au niveau des variables de taille de population et de logements est effectuée par succursale.

Par ailleurs, l'ajout de ces informations par client permet d'obtenir des indicateurs complémentaires, notamment des moyennes, des médianes et des écarts-types pour les variables de population et de logements. De plus, d'autres indicateurs, tels que la proportion de clients résidant dans l'aire de diffusion associée à une succursale, sont ajoutés. Finalement, différentes proportions de clients résidant dans l'un ou l'autre des centres de population ou régions rurales permettent de compléter l'information. Ainsi, pour chaque succursale, les aires de diffusions qui lui sont associées sont transposées en colonnes, afin d'obtenir une ligne par succursale. Mentionnons qu'une succursale peut être associée à jusqu'à 11 069 aires de diffusions. Ensuite, ce fichier de données par succursale est fusionné aux données par clients afin de déterminer si l'une des aires de diffusion de la succursale est la même que celle du client. Cette fusion impliquant des données très volumineuses, le langage *HASH* dans *SAS* a été utilisé, ce qui a permis de mener à bien cette étape dans un délai acceptable.

4.1.3 Données de défavorisation matérielle et sociale pour la santé publique

L'utilisation de données créées pour les besoins de la santé publique dans le cadre d'analyses de performance dans le secteur financier peut être surprenante à première vue. En effet, les différentes pratiques en exploitation de données sont trop peu partagées entre les deux secteurs d'activité. Or, plusieurs similitudes peuvent être constatées. D'abord, alors qu'on peut s'intéresser aux différences dans les pratiques des hôpitaux sur le plan des soins des patients sur un territoire donné dans le domaine de la santé publique par l'utilisation d'indices de défavorisation sociale ou économique, ces mêmes indices peuvent servir à expliquer les différences de performance entre des succursales bancaires étant donné le profil de leurs clients. Ensuite, comme ces indices ont été construits dans le but de compléter les informations présentes dans les bases de données administratives habituellement utilisées dans les études de santé publique, le format des fichiers de données se prête très bien à une intégration aux données habituellement utilisées dans les entreprises bancaires. En effet, les indices sont disponibles par aire de diffusion, une clé de fusion présente également dans les données par succursale du cas d'application considéré. Toutefois, comme les indices sont disponibles depuis quelques années déjà et qu'ils ne semblent pas avoir été mis à jour avec les données de recensement 2011 au moment d'écrire ces lignes, l'aire de diffusion utilisée dans les deux sources de données est celle de 2006. Il semble que l'indice ne soit pas calculé pour environ 2% de la population (Pampalon *et al.*, 2012). Cette proportion de données manquantes a été jugée non problématique, surtout dans un cas d'analyse de données agrégées.

Les données de défavorisation sont ajoutées aux données par client, afin de pouvoir obtenir les moyennes, les médianes et les écarts-types de tous les indicateurs, en fonction du lieu de résidence de chacun des clients d'une succursale donnée. Ainsi, le langage

HASH dans *SAS* a été utilisé pour réduire le temps de traitement requis et pour limiter les besoins en ressources informatiques.

4.1.4 Données de prestation de service

L'intégration des données de prestation de service aux données se fait sur la base de l'identifiant du client et de la succursale. Cette intégration se fait directement dans *SAS* à l'aide de procédures de base et de l'étape *DATA*.

4.1.5 Données de sollicitation et marketing direct

Les informations d'adhésion des succursales aux différentes offensives ne sont pas disponibles dans une base de données structurée, mais plutôt dans des fichiers *Excel* de formats variés, conservées par les ressources humaines qui y ont travaillé. De plus, les types de clients ciblés ne sont pas indiqués dans tous les cas, ce qui requiert un travail de bonification de l'information et de nettoyage des données. Ainsi, des variables ont été développées afin que les différentes informations dans le temps soient toujours présentées de la même façon. Sur toute la période étudiée, les informations de 12 interventions ont été jumelées.

Au total, huit types de clients pouvaient être ciblés dans le temps. Toutefois, ces huit types pouvaient se regrouper en trois grandes cibles. Ce sont les informations relatives à ces trois grandes cibles qui ont été considérées pour la construction de variables finales, d'une part afin de limiter un biais possible lié au raffinement des cibles dans le temps et d'autre part en raison des différents niveaux de précisions utilisés lorsque les informations ont été colligées dans les fichiers de données brutes.

Comme les données finales sont compilées sur une base trimestrielle, la possibilité offerte à la succursale d'exposer ses clients appartenant aux trois cibles a été rapportée dans trois variables prenant la valeur 1 si des clients pouvaient être exposés à la cible pendant ce trimestre et 0 sinon, sur la base de l'offre faite à la succursale. Ensuite, des variables ont été construites pour chacune des cibles afin de considérer si la succursale a accepté ou refusé d'adhérer.

Le nettoyage initial des données a été effectué à l'aide du logiciel *Excel*. Toutefois, le jumelage des différents fichiers *Excel* et la construction des variables finales se sont faits avec le logiciel *SAS*. Un fichier de format « .sas7bdat » a été construit par trimestre, comprenant une observation par succursale. Ces fichiers ont été fusionnés aux fichiers de données obtenus lors des étapes précédentes, sur la base des trimestres correspondants.

4.1.6 Données du site web

Le site web de l'entreprise constitue une source de données complémentaires aux données internes traditionnelles. Toutefois, les données de cette source sont rarement intégrées aux autres données, mais sont plutôt analysées en silo à l'aide de l'outil *Google Analytics*, dans l'optique de valider l'effet de certaines offensives publicitaires, des développements effectués sur le site, certains points relatifs à l'ergonomie ou simplement le niveau de l'achalandage global. La difficulté de trouver une clé de fusion suffisamment fiable est assurément une raison importante liée à l'absence d'intégration de ce type de données. Certaines solutions seraient intéressantes du point de vue de l'utilisation des données, mais ne sont pas implantées ou disponibles au moment d'écrire ces lignes, notamment :

- associer les adresses IP (*Internet Protocol*) des visiteurs du site avec l'identifiant du client lorsque la visite conduit à une connexion dans une zone sécurisée;

- effectuer des requêtes à partir d'un outil payant sur les données brutes pour obtenir des données non agrégées qui permettraient de faire de fins regroupements avec, notamment, des informations géographiques.

En l'absence de ces possibilités, une méthode a été expérimentée avec le nom de la ville dans le cadre de ce cas d'application. En effet, le niveau d'agrégation lié à la position géographique le plus fin disponible pour une extraction à partir de l'outil *Google Analytics* est la ville. Toutefois, en utilisant une méthode d'appariement par le nom de la ville, un biais est causé par les individus visitant le site web de l'entreprise alors qu'ils ne sont pas sur le territoire de leur succursale, par exemple au travail.

La version *Premium* de l'outil *Google Analytics* étant utilisée dans l'entreprise, il a été possible de sortir les données non échantillonnées, à partir de la fonctionnalité de création de rapports personnalisés de l'outil. Comme nous nous intéressons à certaines activités relatives au volet épargne et placement de l'entreprise, un filtre a été utilisé dans la requête afin de ne considérer que les données qui impliquent les pages d'épargne et placement du site web. Ainsi, les visites considérées dans les données web sont celles pour lesquelles au moins une page de la section épargne et placement de l'entreprise a été vue. Afin d'avoir le niveau d'agrégation par ville et heure, ces deux niveaux ont été spécifiés comme dimensions dans l'outil de création de rapports personnalisés.

Les données web obtenues à partir de l'outil *Google Analytics* sont disponibles pour toute la période d'analyse, soit d'octobre 2012 à juin 2014. Toutefois, d'importants changements concernant la mesure des résultats ont été apportés en septembre 2013, ce

qui est fort susceptible d'influencer les résultats d'analyses et les modèles statistiques. Parmi ces changements, mentionnons la migration de l'outil d'analyse de *CoreMetrix* vers *Google Analytics*, de nouvelles définitions pour plusieurs indicateurs, une nouvelle division des pages du site de l'entreprise et un changement dans certaines règles d'affaires pour les filtres d'analyses, notamment la section sur l'épargne et placement. Lors de la migration de l'outil, l'historique de données a toutefois été chargé dans *Google Analytics*, ce qui explique qu'un seul outil d'extraction est utilisé pour ce cas d'application.

Les visites des nombreux employés de l'organisation qui, dans le cadre de leur travail, consultent le site web de l'entreprise pourraient causer des biais non négligeables dans les résultats. Ainsi, afin d'éviter de devoir composer avec un tel problème, toutes les adresses IP des ordinateurs professionnels des employés de l'entreprise ont été considérées comme critère d'exclusion lors de l'extraction des données.

La notion de visiteurs uniques aurait été très intéressante dans les analyses, mais comme les spécialistes de l'analytique web de l'entreprise ne jugent pas l'indicateur valide, cette information n'a pas été retenue pour l'extraction de données. En effet, il semble que l'indicateur de visiteur unique utilisé de pair avec un filtre sur certaines pages cause un problème. Ce problème pourrait être résolu dans le cas où des groupes de contenus seraient disponibles sur les URL (*Uniform Resource Locator*) ciblées par le filtre depuis longtemps, ce qui n'est pas le cas étant donné la migration récente vers l'outil *Google Analytics* pour l'analyse des données web de l'entreprise.

Fusionner des données sur la base d'une clé composée de noms de villes représente un défi de taille, puisqu'un même nom de ville peut être écrit de plusieurs façons différentes. En combinant les données obtenues sur la période des sept trimestres, 4 504 noms de villes différents ont été obtenus. Deux tentatives différentes de fusion ont été testées, mais une seule de ces deux solutions a donné des résultats satisfaisants.

Solution 1 : Transformation de l'orthographe des noms de ville

D'une part, lors de l'ajout des données de Statistique Canada aux données internes traditionnelles, cinq champs pouvant représenter le nom de la ville ont été intégrés. Ces champs sont les suivants :

- nom de la circonscription électorale fédérale (CEFnom);
- nom de la subdivision de recensement (SDRnom);
- nom de la subdivision de recensement unifiée (SRUnom);
- nom de la région métropolitaine de recensement ou de l'agglomération de recensement (RMRnom);
- nom du centre de population et région rurale (CTRPOPRRidu).

D'autre part, un nom de ville est obtenu lors de l'extraction des données de *Google Analytics*. Il s'agit du niveau d'agrégation des données.

Certains éléments dans un champ alphanumérique sont plus susceptibles de causer des problèmes lors de pairage d'observations. Ainsi, un algorithme de transformation du nom de la ville a été appliqué, autant au niveau des noms de villes obtenus lors de l'intégration des données de Statistique Canada que de ceux obtenus à partir de l'outil *Google Analytics*. Ces éléments sont classés en deux catégories : les éléments supprimés et les éléments modifiés. Le Tableau 3 résume les modifications apportées.

Tableau 3 : Classification des éléments modifiés dans les noms de villes

Éléments supprimés	Éléments modifiés	
	Avant	**Après**
espaces traits-d'union (-) apostrophe (‘) virgule (,) point (.) barre oblique (/)	lettres minuscules lettres accentuées (ÉÌÎÎÈÊÀÉÈÔÖÖËÀÏÊËÇÂÏÜÜ)	lettres majuscules lettres non accentuées (EIIIEEAEEOOOEAIEECAIUU)

Toutes ces modifications ont été effectuées en *SAS* à l'aide des fonctions *upcase*, *compress* et *translate*.

Comme cinq champs dans les données de Statistique Canada pouvaient permettre une fusion avec le nom de ville obtenu de *Google Analytics*, la fusion a été tentée cinq fois, en excluant à chaque étape les observations ayant fusionné avec succès à l'étape précédente. À chaque étape, un certain nombre d'observations (villes de *Google Analytics*) ont pu être fusionnées. Voici les résultats obtenus à chacune des étapes :

- étape 1 (SDRnom): 725 villes fusionnées (16,1%);
- étape 2 (SRUnom): 0 ville fusionnée (0,0%);
- étape 3 (CTRPOPRRnom): 84 villes fusionnées (1,9%);
- étape 4 (CEFnom): 3 villes fusionnées (0,1%);
- étape 5 (RMRnom) : 1 ville fusionnée (0,0%).

À la suite de ces cinq étapes, seulement 813 villes de *Google Analytics* (18,1%) sur un total de 4 504 ont été fusionnées, ce qui n'a pas été jugé satisfaisant. La première solution n'a donc pas été retenue pour le jumelage des données.

Solution 2 : Obtention des coordonnées géographiques des villes de *Google Analytics*

Une interface de programmation (en anglais : *Application Programming Interface*, API) rendue disponible par *Google* permet, à partir d'un lieu géographique, d'obtenir les coordonnées géographiques en format latitude et longitude. Cette API, nommée *Google Geocoding*, a été utilisée par l'organisation *Police Analyst*[6] dans la création d'une fonction *Excel* permettant d'obtenir les latitudes et longitudes de lieux géographiques. Cette fonction est accessible sur le web par le site de l'organisation.[7] Cette fonction a donc été utilisée pour convertir les noms de villes *Google Analytics* en coordonnées de latitude et longitude, après avoir suivi les différentes étapes précisées sur la page web. Quelques ajustements ont toutefois été nécessaires, puisque l'utilisation de la fonction intacte engendrait des erreurs dans *Excel*. Ces erreurs étaient probablement dues à l'utilisation d'une version ultérieure du logiciel que celle mentionnée dans le tutoriel de *Police Analyst* (*Excel* 2013 versus *Excel* 2007). Les différentes étapes pour l'utilisation de cette fonction avec *Excel* 2013 sont présentées en Annexe 3.

Cette fonction est facile à utiliser et ne nécessite pas de références à des adresses complètes. Ainsi, le nom de la ville, dans plusieurs cas, est suffisant. Dans une feuille *Excel*, la syntaxe utilisée est la suivante :

=googlegeocode(cellule_contenant_nom_de_ville)

---

[6] http://policeanalyst.com/
[7] http://policeanalyst.com/using-the-google-geocoding-api-in-excel/,

Or, en expérimentant l'utilisation de cette fonction sur les 4 504 villes *Google Analytics* distinctes obtenues de l'extraction, des difficultés sont rapidement constatées. Tout d'abord, l'utilisation de cette fonction sur un grand nombre de villes nécessite beaucoup trop de ressources de la part de l'ordinateur. De plus, une limite journalière de 2 500 requêtes par 24 heures pour les utilisateurs de la version gratuite est imposée par *Google*. Finalement, si l'ordinateur est relativement puissant et est en mesure de traiter un certain nombre de requêtes simultanément, le temps de réponse est trop rapide et plusieurs données sont retournées manquantes par *Google*. En effet, *Google* limite le nombre de requêtes à 10 par secondes. Ainsi, en considérant ces contraintes, une procédure permettant l'obtention des latitudes et longitudes a été développée et est présentée en Annexe 4.

Au total, 3 525 noms de villes ont pu être associés à des coordonnées géographiques dès la première étape de requêtes ou après une soumission manuelle de la fonction sans ajouts de précisions. Ensuite, 895 noms de villes ont été associés à des coordonnées géographiques après l'ajout d'une précision géographique (Québec, Ontario, Canada, US). Finalement, 84 noms de villes n'ont pas été associés à des coordonnées géographiques et ont été exclus de la liste. En résumé, 98% des noms de villes de *Google Analytics* ont été associés à des coordonnées géographiques.

Une fois les coordonnées géographiques obtenues pour les villes *Google Analytics*, toutes les tables contenant les noms de villes sont importées dans *SAS*, afin de former un seul fichier *SAS*. C'est d'ailleurs dans *SAS* que s'effectuent les prochaines étapes.

Afin de déterminer les villes devant être associées à chacune des succursales, une jointure en produit cartésien est effectuée. Ainsi, toutes les possibilités de villes *Google Analytics* avec chaque succursale sont évaluées, en calculant la distance entre les coordonnées géographiques de latitude et longitude de la succursale (obtenues à partir

de l'aire de diffusion et des données de Statistique Canada) et les coordonnées géographiques des villes *Google Analytics*. Cette distance est obtenue en utilisant la fonction « geodist » de *SAS*. Mentionnons que cette fonction permet d'obtenir la distance à vol d'oiseau, ce qui engendre des problématiques au niveau de villes de part et d'autre d'obstacles naturels tels que des cours d'eau (le fleuve St-Laurent en est un exemple), mais que ce type de problématique est ignoré dans le cadre de ce projet.

Pour déterminer les villes devant être associées à chaque succursale, la mesure de l'écart-type de la distance entre l'aire de diffusion du client et celle de sa succursale a été jugée utile, puisqu'elle permet de moduler la distance maximale à considérer en fonction du contexte d'étalement géographique de la clientèle autour de la succursale. En effet, il aurait été inapproprié de fixer une distance maximale, égale pour chaque succursale, puisque les territoires peuvent être très différents, notamment lorsque des succursales en région urbaine et en région rurale sont comparées. Ainsi, toutes les villes *Google Analytics* ayant une distance plus petite ou égale à la médiane plus l'écart-type (calculé par rapport à la moyenne) de la distance entre la succursale et l'aire de diffusion du client sont conservées pour la succursale. La médiane est privilégiée par rapport à la moyenne afin d'éviter de tenir compte de distances extrêmes entre les résidences des clients et leur succursale. Mentionnons que toutes les succursales sont associées à au moins une ville *Google Analytics* et que plusieurs succursales peuvent être associées à une même ville *Google Analytics*. En considérant les données trimestrielles des villes *Google Analytics*, le Tableau 4 présente certaines statistiques concernant le nombre de villes *Google Analytics* associées aux succursales, en fonction de la méthode de traitement décrit. Les petits villages sont considérés comme étant des villes pour *Google Analytics*, ce qui justifie le nombre élevé de villes.

Tableau 4 : Statistiques sur l'assignation des villes de *Google Analytics* aux succursales (en nombre de villes par succursale)

	Nombre de villes par succursale					Pondération des résultats de la ville GA associée à la première succursale en importance (en %)			
Année et Trimestre	**Min**	**Max**	**Moyenne**	**Écart-type**	**Médiane**	**Moyenne**	**Q1**	**Q2 (médiane)**	**Q3**
2012 T4	1	475	41,79	43,58	30,00	57,20	16,18	57,10	100,00
2013 T1	1	580	42,49	47,52	30,00	57,24	18,78	62,99	100,00
2013 T2	1	474	40,21	41,31	30,00	55,91	16,45	54,64	100,00
2013 T3	1	571	46,22	47,57	34,00	58,15	18,31	60,05	100,00
2013 T4	1	480	42,48	42,52	29,50	56,62	16,78	55,81	100,00
2014 T1	1	670	49,38	53,37	36,00	60,20	18,85	64,41	100,00
2014 T2	1	592	47,63	49,83	33,00	59,00	18,61	63,67	100,00

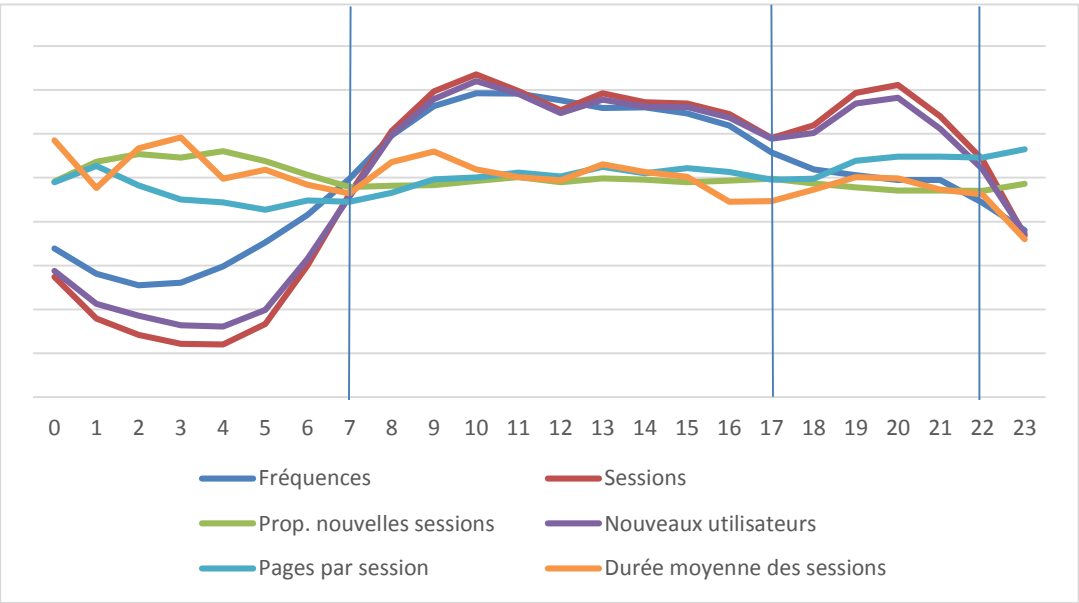
Comme une ville *Google Analytics* peut être associée à plus d'une succursale, il est nécessaire de répartir les résultats trimestriels de cette ville entre les succursales, afin d'éviter d'utiliser la même information plus d'une fois. Toutefois, l'hypothèse que les résultats de la ville *Google Analytics* n'ont pas le même apport pour chacune des succursales à laquelle elle est associée est posée. Ainsi, afin de déterminer la pondération devant être associée à chaque succursale pour les résultats d'une ville *Google Analytics*, une fonction utilisant l'inverse de la distance entre la succursale et la ville *Google Analytics* est utilisée. De cette façon, les succursales plus près de la ville prennent une plus grande proportion des résultats que les succursales plus éloignées, en respectant le critère de l'écart-type. Le calcul de la pondération de la succursale pour la ville *Google Analytics* se fait donc de la façon suivante :

$$w_i = \frac{{d_i}^{-1}}{\sum_{i=1}^{N} {d_i}^{-1}}$$

où chaque *i* représente une succursale associée à la ville *Google Analytics* et *d* représente la distance entre la succursale et la ville. Mentionnons que pour chaque ville *Google Analytics*, la somme des poids est égale à un.

Les données obtenues de *Google Analytics* ont été extraites pour chaque heure de la journée, par ville. De cette façon, il est possible de tenir compte du fait que le comportement des clients sur le web peut varier en fonction des moments de la journée, ce qui permet d'apporter des informations supplémentaires. Toutefois, afin de ne pas créer 24 variables pour chaque indicateur, l'évolution moyenne des valeurs des indicateurs pendant la journée est analysée graphiquement. Pour ce faire, la moyenne des résultats de toutes les villes a d'abord été calculée pour chacun des indicateurs, par heure. Ensuite, afin de visualiser toutes les courbes sur une échelle comparable, l'écart des résultats de chaque heure par rapport à la moyenne de toutes les heures pour tous les indicateurs a été calculé. Les résultats obtenus sont présentés à la Figure 3.

Figure 3 : Graphique présentant l'évolution moyenne relative des indicateurs web, par heure



À l'observation de ce graphique, les différentes heures de la journée sont regroupées en trois catégories en fonction de l'évolution des valeurs des indicateurs. Ces subdivisions sont placées en fonction d'un critère visuel et du contexte des périodes d'activités habituelles d'une journée. Ainsi, les données mesurées entre 22 heures et 7 heures forment la première catégorie, celles mesurées entre 7 heures et 17 heures constituent la seconde catégorie et celles mesurées entre 17 heures et 22 heures sont regroupées dans la troisième catégorie. Les six indicateurs obtenus de *Google Analytics* sont donc représentés dans 18 variables (trois variables par indicateurs).

Une fois toutes les manipulations effectuées sur les données provenant de l'outil *Google Analytics*, l'ajout de ces données aux données des autres sources se fait en fusionnant les résultats agrégés par succursale.

4.2 Variables : types et construction

Les variables initiales obtenues à partir des différentes sources constituent le bassin total des informations pertinentes à la construction des variables finales qui sont utilisées dans les modèles. En effet, d'une part, le fait d'analyser des données agrégées nécessite un travail important de construction de variables afin de conserver l'information totale contenue dans les différents types de variables initiales, tel que les variables qualitatives de type nominales ou ordinales. D'autre part, il est intéressant de considérer la création de nouvelles variables issues de transformations. Cette étape « n'est pas systématique et certains modèles n'utilisent que les variables initiales, mais elle présente souvent de l'intérêt dans les problématiques non scientifiques, où les variables n'ont pas été collectées en vue d'une analyse statistique et ne sont pas forcément les mieux adaptées à la problématique » (Tufféry, 2010). Ce dernier énoncé qualifie justement très bien la situation particulière de notre cas d'application, puisque les données n'ont justement pas été collectées en vue de cette étude.

Les sections suivantes font état du processus de réflexion relatif à la sélection des variables finales.

4.2.1 Variable dépendante (à expliquer)

Le suivi des résultats de ventes nettes est une activité importante dans l'entreprise. La vente nette constitue un indicateur de choix, puisqu'elle permet de considérer le développement des affaires et la rétention de la clientèle dans une seule mesure. Ainsi, il est très pertinent de considérer les ventes nettes dans les modèles statistiques, en tant que variable dépendante (à expliquer). Or, afin de comparer toutes les succursales sur une base similaire, la vente nette sera une composante d'un ratio. Ce ratio sera le suivant :

Ventes nettes succursale / Volume d'affaires de la succursale

4.2.2 Variables indépendantes (explicatives)

Les variables explicatives (variables indépendantes) ont été choisies en fonction de ce que les différentes sources de données rendaient disponible et afin de permettre une vue complète de la réalité d'une succursale dans un contexte d'affaires dans le domaine de l'épargne et placement. Les variables sélectionnées peuvent être classées dans l'une ou l'autre de ces catégories :

- variables géographiques;
- variables financières;
- variables sociodémographiques;
- variables comportement des clients;
- variables prestation de service et sollicitation;
- variables site web.

Au total, 406 variables ont été choisies pour leur potentiel explicatif de la performance des ventes en épargne dans les succursales. Une fois les variables de référence retirées, ce nombre est de 379. Ces variables ont été sélectionnées sur la base de la connaissance du milieu, du contexte d'affaires et suite à des échanges avec des professionnels et gestionnaires de l'entreprise. En Annexe 5, un tableau présente la liste complète des variables construites à partir des données obtenues des différentes sources.

La plupart des variables proviennent de la manipulation ou de la transformation des données extraites des différentes sources de données. Plusieurs types de manipulations sont effectués en fonction de la nature de la donnée et de la façon dont elle sera utilisée. De plus, étant donné le format agrégé par succursale des données finales, il importe de sélectionner des types de manipulations qui permettent de faire parler les données efficacement. Ces manipulations peuvent être des sommes, des proportions, des moyennes, des écarts-types et des ratios.

Les sommes

Les sommes sont très peu utilisées pour les variables finales, puisqu'elles ne permettent pas de comparer les succursales entre elles, les succursales pouvant être de tailles différentes. La seule somme utilisée concerne le volume d'affaires de la succursale, puisqu'il témoigne de sa taille et que cette variable pourrait permettre d'expliquer en partie sa performance relative.

Les proportions

Les données initiales peuvent parfois être des données de type catégorique. Ainsi, avant d'agréger les données, il importe de construire des variables distinctes pour chacune des catégories, qui prennent la valeur 1 si l'observation appartient à cette catégorie et 0 sinon. Ces données catégoriques touchent principalement les données de type sociodémographique ou géographique. Lors de l'agrégation, les proportions des observations associées à la succursale appartenant à l'une ou l'autre des catégories sont calculées. Des proportions de données manquantes sont également calculées et considérées à part. Les proportions permettent une comparaison des succursales entre elles, malgré leurs tailles différentes.

Les moyennes et les écarts-types

Pour toutes les variables continues, des moyennes et des écarts-types par succursale sont calculés. D'une part, les moyennes permettent de comparer les succursales entre elles, et ce, peu importe leur taille. Elles peuvent permettre d'avoir le portrait moyen d'un client de la succursale. D'autre part, les écarts-types permettent de constater la variabilité de certains indicateurs à l'intérieur d'une succursale. Cette variabilité pourrait être un élément explicatif de sa performance. Les moyennes et les écarts-types sont principalement associés aux variables financières.

Les médianes

Les médianes sont construites afin de répondre à un besoin similaire que pour les moyennes, à l'exception du fait que, dans ce cas, elles sont choisies afin d'éliminer l'impact que pourraient avoir des données extrêmes.

Les ratios

Les ratios permettent de mettre en relation deux éléments qui ont un meilleur pouvoir explicatif lorsqu'ils sont regroupés que lorsqu'ils sont considérés séparément. Ils permettent également de mettre en valeur le poids d'un élément par rapport à un autre. Ils sont principalement utilisés dans le cas de variables financières. Comme pour la proportion et la moyenne, le ratio est une mesure stable qui permet la comparaison des succursales entre elles.

Chapitre 5 : Analyse des résultats

5.1 Résultats de la sélection des modèles

Étant donné l'objectif de cette recherche de vérifier si l'ajout de données de sources externes ou inhabituelles aux données des entrepôts internes d'une entreprise du secteur bancaire permet d'améliorer la performance de modèles statistiques, il importe de faire la comparaison de modèles construit avec des données d'entrepôts de données internes uniquement (modèle 1) et construit avec des variables provenant de sources multiples de données (modèle 2).

5.1.1 Utilisation des données des entrepôts internes de l'entreprise (modèle 1)

255 variables sont considérées au départ pour la construction du modèle à partir des données des entrepôts internes de l'entreprise. Au terme de l'étape de présélection de variables, 43 variables sont conservées. Ces variables sont de plusieurs types et sont présentées dans le Tableau 5.

Tableau 5 : Variables conservées après l'étape de présélection, modèle 1

Variables géographiques
- Nombre moyen de succursales du concurrent 3 dans l'aire de diffusion du client
- Nombre moyen de succursales du concurrent 6 dans l'aire de diffusion du client
Variables financières
- Montant total de ventes nettes en produits virtuels
- Montant médian d'écart entre l'épargne estimée totale dans le marché et l'épargne détenue dans l'entreprise
- Montant médian autorisé pour l'ensemble des prêts
- Montant moyen d'écart entre l'épargne estimée totale dans le marché et l'épargne détenue dans l'entreprise
- Montant moyen d'épargne détenue dans l'entreprise
- Montant moyen d'épargne estimée totale dans le marché
- Proportion de clients détenant un produit d'épargne 3
- Proportion de clients détenant un produit d'épargne 6
- Proportion de clients faisant affaire avec la filiale de valeurs mobilières
- Proportion de clients détenteurs d'une carte de crédit pour « Particuliers » de l'entreprise
- Proportion de clients détenant un produit d'épargne 1
- Montant moyen du chiffre d'affaires des entreprises
- Montant moyen détenu dans le produit d'épargne 9
- Montant médian de revenu brut
- Ratio montant d'écart d'épargne estimée totale marché / montant total estimé marché
- Ratio montant d'épargne détenue dans l'entreprise / montant total estimé marché
- Écart-type du montant détenu dans le produit d'épargne 4
- Écart-type du montant détenu dans le produit d'épargne 8
Variables sociodémographiques
- Proportion de clients entreprises légales de type Ltd
- Âge médian
- Proportion de clients dont l'état civil est « initial »
- Proportion de clients dont l'état civil est « célibataire »
- Proportion de clients dont l'état civil est « marié »
- Proportion de clients dont l'état civil est « conjoint de fait »
- Proportion de clients – cycle de vie « accédant à la propriété »
- Proportion de clients – cycle de vie « nouveaux propriétaires »

- Proportion de clients – cycle de vie « retraite »
- Proportion de clients démunis
- Proportion de clients appartenant à une communauté culturelle
- Proportion de clients dont le niveau de richesse est « bâtisseur »
- Proportion de clients dont le niveau de richesse est « Aisé 1 »
- Proportion de clients dont le code d'occupation est « profession libérale ou technicien »
- Proportion de clients dont le code d'occupation est « ouvrier spécialisé »
- Proportion des clients locataires
- Proportion des clients chambreurs
- Proportion des clients habitant chez leurs parents
- Proportion d'entreprises d'exploitation immobilière résidentielle
Variables comportement
- Proportion des clients dont la relation d'affaires est « mixte financement »
- Proportion des clients utilisant les services par téléphone pour de l'information seulement
- Proportion des clients utilisant les services par téléphone pour leurs transactions
Autres variables
- Temps

Une fois cette présélection effectuée, il importe de détecter la structure de covariance la plus appropriée pour ces données. Pour ce faire, des modèles sont ajustés avec l'ensemble des variables présélectionnées à l'étape précédente, pour 9 structures de covariance différentes. Ces modèles tiennent compte des mesures répétées relatives à chacune des succursales. Le Tableau 6 présente les résultats obtenus sur la base des critères de sélection *AIC* et *BIC*, pour chacune des structures de covariances.

Tableau 6 : Comparaison de modèles saturés en fonction de la structure de covariance, modèle 1

Structure	AIC	BIC
Aucune corrélation intrasuccursale	-6509,4	-6503,6
AR(1)	-6689,5	-6681,8
CS	-6579,2	-6571,4
TOEP	-7629,6	-7602,4
VC	-6509,4	-6505,5
ARMA(1,1)	-7131,2	-7119,5
UN	-12770,0	-12661,1
ARH(1),	-8975,0	-8943,9
ANTE(1)	-12056,2	-12005,7
HF	-6833,9	-6802,6

Les critères de sélection *AIC* et *BIC* permettent de faire certains constats. En premier lieu, tous les modèles pour lesquels une structure de covariance a été spécifiée performent mieux que lorsque le contexte de mesures répétées n'est pas pris en compte. Il est donc pertinent de tenir compte d'une certaine corrélation entre les observations pour une succursale. Ensuite, le choix de la structure de covariance *non structurée* se distingue favorablement des autres choix et a pour caractéristique de ne pas tenir compte d'une structure a priori entre les mesures. L'analyse de la matrice de corrélation du ratio de performance (variable dépendante) pour une succursale donnée permet de mieux comprendre le lien unissant les différentes observations entre elles. La matrice de corrélation d'une succursale est présentée à la Figure 4. Ainsi, voici quelques éléments intéressants par rapport au contexte du cas d'application et des données qui sont analysées :

- le nombre de trimestres d'écart entre deux mesures est déterminant dans l'évaluation de la corrélation. Plus les trimestres sont rapprochés, plus la corrélation, positive ou négative, est forte. Cette corrélation tend à diminuer lorsque le nombre de trimestres d'écart augmente;

- Le trimestre 4 d'une année est très fortement corrélé avec le trimestre 1 de l'année suivante;

- Le trimestre 2 est très fortement corrélé avec le trimestre 3 qui suit;

- d'une année à l'autre, les observations suivent des dynamiques de corrélation similaires;

- Les trimestres 4 de 2012 et 1 de 2013 sont inversement corrélés avec le reste de l'année 2013 et le début de 2014;

- le trimestre 2 de 2014 constitue une exception. Il est très peu corrélé avec les autres trimestres.

Figure 4 : Estimation de la matrice de corrélation pour une succursale, structure de covariance *non spécifiée*, modèle saturé à partir de données des entrepôts de données internes

		2012	2013				2014	
		T4	T1	T2	T3	T4	T1	T2
2012	T4	1,00	0,97	-0,72	-0,76	-0,32	-0,43	0,06
2013	T1	0,97	1,00	-0,85	-0,88	-0,47	-0,56	0,00
	T2	-0,72	-0,85	1,00	0,99	0,72	0,76	0,24
	T3	-0,76	-0,88	0,99	1,00	0,73	0,76	0,22
	T4	-0,32	-0,47	0,72	0,73	1,00	0,91	0,04
2014	T1	-0,43	-0,56	0,76	0,76	0,91	1,00	0,11
	T2	0,06	0,00	0,24	0,22	0,04	0,11	1,00

En utilisant la structure de covariance *non structurée* (*UN*), la méthodologie de sélection des effets fixes (étape 3) est appliquée. Le processus est arrêté au moment où le « meilleur » modèle comprenant 22 effets fixes est ajusté, puisque les valeurs des critères *AIC* et *BIC* ont atteint leurs minimums quelques étapes auparavant (détails du processus de sélection en Annexe 6). Le critère *BIC* le plus faible favorise un modèle à 14 effets fixes, alors que le critère *AIC* favorise un modèle un peu plus complexe avec 18 effets fixes. Les variables sélectionnées en considérant les deux critères de sélection sont présentées à la Figure 5. Des variables de nature temporelle, financière, sociodémographique et comportementale sont sélectionnées autant sur la base du critère *BIC* que du critère *AIC*.

Figure 5 : Variables sélectionnées dans les modèles favorisés par les critères *AIC* et *BIC*, utilisation des données des entrepôts internes

Variables	*AIC*	*BIC*
Temps	X	X
Montant total de ventes nettes en produits virtuels	X	X
Montant médian autorisé pour l'ensemble des prêts	X	X
Proportion de clients dont l'état civil est « marié »	X	X
Proportion de clients appartenant à une communauté culturelle	X	X
Proportion des clients locataires	X	X
Écart-type du montant détenu dans le produit d'épargne 4	X	X
Proportion de clients détenant un produit d'épargne 3	X	X
Proportion des clients dont la relation d'affaires est « mixte financement »	X	X
Proportion de clients dont le code d'occupation est « profession libérale ou technicien »	X	X
Proportion de clients dont l'état civil est « célibataire »	X	X
Proportion de clients dont l'état civil est « initial »	X	X
Proportion des clients habitant chez leurs parents	X	X
Ratio montant d'écart d'épargne estimée totale marché / montant total estimé marché	X	X
Proportion de clients détenant un produit d'épargne 1	X	
Écart-type du montant détenu dans le produit d'épargne 8	X	
Proportion de clients dont le niveau de richesse est « Aisé 1 »	X	
Proportion des clients chambreurs	X	

Étant donné la nature variée des variables sélectionnées sur la base des deux critères, le modèle plus parcimonieux sera retenu pour la suite, soit celui associé à la plus faible valeur de *BIC*. Ainsi, certaines validations sur le modèle comprenant 14 variables seront effectuées.

Tout d'abord, l'analyse du graphique des résidus en fonction des valeurs prédites permet de constater que les résidus sont distribués aléatoirement autour de zéro et que quelques prédictions sont négatives. De plus, le graphique présentant les résidus en fonction des quantiles de la loi normale montre une ligne droite sur une importante partie de la courbe, à l'exception d'une quinzaine de points qui contribuent à faire courber l'extrémité gauche et une vingtaine de points qui contribuent à faire courber l'extrémité droite. Par rapport au nombre total d'observations, ce nombre de points est très petit, ce qui porte à conclure qu'il est plausible que la distribution des erreurs soit suffisamment près d'une loi normale pour se fier aux valeurs des paramètres estimés et aux seuils observés obtenus par le modèle. Les graphiques analysés sont présentés en Annexe 8.

Ensuite, l'observation des graphiques de résidus en fonction des variables indépendantes permet de détecter une présence potentielle d'hétéroscédasticité pour trois variables, soient :

- l'écart-type du montant détenu dans le produit d'épargne 4;
- le montant total de ventes nettes en produits virtuels;
- la proportion de clients appartenant à une communauté culturelle.

Effectivement, la distribution des résidus en forme d'entonnoir sur les graphiques semble suggérer que ce type de problème est présent. Par contre, le seuil observé de 0,1539 du test obtenu avec l'option *SPEC* de *PROC REG* ne permet pas de rejeter l'hypothèse nulle d'homogénéité des variances au seuil de 5 %. Étant donné le problème suspecté par l'observation des graphiques, des transformations sur ces variables sont tout de même tentées, ainsi que des transformations sur la variable dépendante. La transformation « racine carrée » appliquée sur les trois variables explicatives est celle qui donne les meilleurs résultats, tant au niveau de l'amélioration de la distribution des résidus constatée par l'observation des graphiques que sur le résultat du test de l'homogénéité des variances dont le seuil observé passe à 0,2854. Finalement, un diagnostic de la multicolinéarité avec l'utilisation de l'option *COLLINOINT* de la procédure *REG* est

opéré. Comme les valeurs de l'indicateur « *Condition Index* » du tableau « *Collinearity Diagnostics (intercept adjusted)* » sont toutes inférieures à 5,5, aucun problème de multicolinéarité ne semble présent dans ce modèle. De plus, toutes les valeurs de l'indicateur de l'inflation de la variance associées aux variables explicatives sont inférieures à 4, ce qui confirme l'absence d'un problème de multicolinéarité. En guise de complément de validation portant sur la détection d'un problème de multicolinéarité, des modèles univariés avec chacun des effets fixes sélectionnés sont comparés avec le modèle complet, afin de valider si le signe des paramètres estimés est le même dans les deux cas. Presque tous les paramètres estimés sont de même signe, à l'exception de ceux relatifs aux variables « Proportion des clients locataires » et « Proportion de clients dont le code d'occupation est « profession libérale ou technicien », qui sont négatifs dans le modèle multivarié, mais positifs dans les modèles univariés. Cette différence peut être due à la présence de multicolinéarité avec d'autres variables. Afin de comprendre qu'elles étaient les variables en cause, le modèle a été ajusté successivement en retirant tour à tour les variables du modèle. Les modèles pénalisés d'une variable auraient pu permettre de constater que le signe des paramètres estimés change lorsque la variable impliquée dans une relation de multicolinéarité est retirée. Or, aucun changement de signe n'a été observé sur tous les paramètres estimés.

Au terme de ces validations et des corrections appropriées, la valeur du critère *BIC* diminue (-13110,1 contre -13105,09 précédemment), ce qui permet de constater que les transformations de variables permettent d'améliorer le modèle.

Intégration de certains termes d'interaction et d'ordre supérieur

Le modèle obtenu après l'étape de validation peut très bien convenir dans un but d'explication des résultats de performance des succursales au niveau des ventes en épargne. Or, pour aller plus loin et améliorer la précision du modèle, l'intégration de termes d'ordre supérieur peut être considérée. En effet, des interactions entre des

variables ou des variables intégrées à une puissance supérieure à un peuvent donner des informations complémentaires que les effets fixes intégrés jusqu'à maintenant ne peuvent fournir. Toutefois, l'ajout de nouveaux effets fixes de ce type complexifie le modèle et rend beaucoup plus difficile son interprétation. Cet ajout serait surtout utile pour une utilisation d'un modèle à des fins de prédictions. L'ajout de termes d'ordre supérieur a tout de même été tenté pour ce modèle, afin de démontrer à quel point le modèle peut être amélioré. Toutefois, cette extension au modèle ne sera pas considérée lors de l'interprétation du modèle dans un but de compréhension des éléments qui affectent la performance des succursales dans le contexte des ventes en épargne. Seuls des termes d'interaction entre deux variables ou des variables au carré sont considérés ici; cette section doit être vue comme étant le début d'un processus de perfectionnement du modèle et non une finalité. Les détails concernant la sélection d'effets fixes d'ordre supérieur sont présentés à l'Annexe 7.

Au terme du processus d'ajouts d'effets fixes qui consiste à identifier les meilleurs modèles avec l'ajout de termes supplémentaires jusqu'à ce que l'augmentation du nombre d'effets ne permette plus de diminuer la valeur du critère *BIC* pendant quelques étapes, un modèle comprenant l'ajout de 17 effets fixes d'ordre supérieur est sélectionné. En effet, la valeur du *BIC* a constamment diminué jusqu'à ce nombre, pour ensuite remonter. L'ajout de ce grand nombre d'effets fixes permet d'améliorer de façon importante le niveau de précision du modèle, puisque la valeur du critère *BIC* selon la méthode *ML* passe de -13110,1 à -13198,8. Ainsi, en considérant le développement d'un tel modèle dans un objectif de prévision, l'ajout d'interactions et de termes d'ordre supérieur constituerait une amélioration importante.

Pour terminer, la structure de covariance a été de nouveau validée en fonction du modèle final, en incluant les interactions et termes d'ordre supérieurs. Cette vérification a permis de constater que la structure *non structurée* est toujours celle qui donne les meilleurs résultats sur le plan de valeurs obtenues au niveau des critères de sélection parmi les différentes structures considérées au départ.

5.1.2 Ajout des autres sources de données (modèle 2)

La méthodologie de sélection de modèles et de variables est la même que celle utilisée pour le modèle intégrant des données des entrepôts de données internes. 376 variables sont considérées au départ pour la construction du modèle intégrant les données provenant des multiples sources. Au terme de l'étape de présélection de variables, 38 variables sont conservées. Ces variables sont de plusieurs catégories et sont présentées dans le tableau 7.

Tableau 7 : Variables conservées après l'étape de présélection, modèle 2

Variables géographiques
- Nombre moyen de succursales du concurrent 5 dans l'aire de diffusion du client
- Nombre moyen de succursales du concurrent 6 dans l'aire de diffusion du client
- Écart-type de nombre de succursales du concurrent 5 dans l'aire de diffusion du client
- Ratio nombre de logements privés occupés / population des aires de diffusion
- Proportion de clients habitant à l'extérieur d'une région métropolitaine, aucune influence métropolitaine
- Distance moyenne entre la résidence principale des clients et la succursale
Variables financières
- Montant total de ventes nettes virtuelles
- Montant médian autorisé pour l'ensemble des prêts
- Proportion de clients détenant un produit d'épargne 3
- Proportion de clients détenant un produit d'épargne 4
- Proportion de clients faisant affaire avec la filiale de gestion discrétionnaire
- Montant moyen détenu dans le produit d'épargne 4
- Montant moyen détenu en gestion discrétionnaire
- Écart-type du montant détenu dans le produit d'épargne 4
- Écart-type du montant détenu en gestion discrétionnaire
- Proportion de clients attitrés à l'équipe 3
Variables sociodémographiques
- Proportion de clients dont l'état civil est « initial »

- Proportion de clients dont l'état civil est « célibataire »
- Proportion de clients dont l'état civil est « séparé »
- Proportion de clients dont l'état civil est « conjoint de fait »
- Proportion de clients – cycle de vie « accédant à la propriété »
- Proportion de clients – cycle de vie « retraite »
- Proportion de clients appartenant à une communauté culturelle
- Proportion de clients dont le niveau de richesse est « bâtisseur »
- Proportion de clients dont le niveau de richesse est « Aisé 1 »
- Proportion de clients dont le code d'occupation est « profession libérale ou technicien »
- Proportion de clients dont le code d'occupation est « vendeur »
- Proportion de clients dont le code d'occupation est « ouvrier spécialisé »
- Proportion des clients locataires
- Proportion des clients chambreurs
- Proportion des clients habitant chez leurs parents
- Proportion d'investisseurs immobiliers
Variables prestation de service et sollicitation
- Offre d'adhésion au plan de sollicitation de la cible E
Variables web
- Durée moyenne des sessions dans le groupe d'heures 2
- Durée moyenne des sessions dans le groupe d'heures 3
- Ratio nombre de sessions dans le groupe d'heures 2 / nombre de sessions totales
- Ratio nombre de sessions dans le groupe d'heures 3 / nombre de sessions totales
Variables comportement
Autres variables
- Temps

Le Tableau 8 présente les résultats obtenus sur la base des critères de sélection *AIC* et *BIC*, pour 9 structures de covariances. Comme pour le modèle de la section précédente, la structure *non structurée (UN)* est associée aux meilleurs résultats considérant les valeurs obtenues pour les critères *AIC* et *BIC*. L'analyse effectuée relativement aux

corrélations entre les différents trimestres présentée dans la section du modèle 1 s'applique également dans le cadre de ce modèle. La matrice de corrélations du ratio de performance (variable dépendante) pour une succursale est présentée à la Figure 6. Dans l'ensemble, les valeurs des corrélations pour cette succursale sont très similaires à celles obtenues pour le modèle construit à partir des données des entrepôts de données internes (Figure 4).

Tableau 8 : Comparaison de modèles saturés en fonction de la structure de covariance, modèle 2

Structure	AIC	BIC
Aucune corrélation intrasuccursale	-6640,1	-6634,2
AR(1)	-6791,3	-6783,5
CS	-6735,4	-6727,7
TOEP	-7686,6	-7659,4
VC	-6640,1	-6636,2
ARMA(1,1)	-7233,3	-7221,6
UN	-12918,6	-12809,8
ARH(1),	-9132,4	-9101,3
ANTE(1)	-12218,7	-12168,2
HF	-6962,2	-6931,1

Figure 6 : Estimation de la matrice de corrélation pour une succursale, structure de covariance *non structurée*, modèle saturé à partir de données de multiples sources (modèle 2)

		2012	2013				2014	
		T4	T1	T2	T3	T4	T1	T2
2012	T4	1,00	0,97	-0,74	-0,78	-0,39	-0,5	0,00
2013	T1	0,97	1,00	-0,86	-0,89	-0,52	-0,62	-0,05
	T2	-0,74	-0,86	1,00	0,99	0,74	0,78	0,28
	T3	-0,78	-0,89	0,99	1,00	0,75	0,78	0,26
	T4	-0,39	-0,52	0,74	0,75	1,00	0,91	0,04
2014	T1	-0,50	-0,62	0,78	0,78	0,91	1,00	0,12
	T2	0,00	-0,05	0,28	0,26	0,04	0,12	1,00

En utilisant la structure de covariance *non structurée*, la méthodologie de sélection des effets fixes (étape 3) est appliquée. Le processus est arrêté au moment où le « meilleur » modèle comprenant 33 effets fixes est ajusté, puisque les valeurs des critères *AIC* et *BIC* ont atteint leurs minimums quelques étapes auparavant (détails du processus de sélection en Annexe 6). Le critère *BIC* le plus faible favorise un modèle à 11 effets fixes, alors que le critère *AIC* favorise un modèle plus complexe avec 30 effets fixes. Les variables sélectionnées en considérant les deux critères de sélection sont présentées à la Figure 7. Des variables de nature temporelle, financière, sociodémographique, web et géographique sont choisies sur la base du critère *BIC*, alors qu'une variable de sollicitation s'ajoute en considérant le modèle sélectionné sur la base du critère *AIC*.

Figure 7 : Variables sélectionnées dans les modèles favorisés par les critères *AIC* et *BIC*, utilisation des données provenant de multiples sources

Variables	*AIC*	*BIC*
Temps	X	X
Montant total de ventes nettes virtuelles	X	X
Durée moyenne des sessions dans le groupe d'heures 2	X	X
Montant médian autorisé pour l'ensemble des prêts	X	X
Proportion de clients dont l'état civil est « initial »	X	X
Proportion de clients dont l'état civil est « célibataire »	X	X
Proportion de clients détenant un produit d'épargne 3	X	X
Proportion de clients appartenant à une communauté culturelle	X	X
Proportion de clients habitant à l'extérieur d'une région métropolitaine, aucune influence métropolitaine	X	X
Écart-type du montant détenu dans le produit d'épargne 4	X	X
Ratio nombre de sessions dans le groupe d'heures 3 / nombre de sessions totales	X	X
Nombre moyen de succursales du concurrent 5 dans l'aire de diffusion du client	X	
Écart-type de nombre de succursales du concurrent 5 dans l'aire de diffusion du client	X	
Proportion des clients locataires	X	
Proportion des clients habitant chez leurs parents	X	
Proportion de clients dont le code d'occupation est « profession libérale ou technicien »	X	
Ratio nombre de logements privés occupés / population des aires de diffusion	X	
Proportion de clients dont le code d'occupation est « vendeur »	X	
Proportion de clients dont l'état civil est « séparé »	X	
Distance moyenne entre la résidence principale des clients et la succursale	X	
Proportion de clients détenant un produit d'épargne 4	X	
Proportion des clients chambreurs	X	

Offre d'adhésion au plan de sollicitation de la cible E	X
Ratio nombre de sessions dans le groupe d'heures 2 / nombre de sessions totales	X
Proportion de clients dont le niveau de richesse est « bâtisseur »	X
Proportion de clients dont l'état civil est « conjoint de fait »	X
Proportion d'investisseurs immobiliers	X
Nombre moyen de succursales du concurrent 6 dans l'aire de diffusion du client	X
Proportion de clients dont le niveau de richesse est « Aisé 1 »	X
Proportion de clients attitrés à l'équipe 3	X

Aux fins de comparaisons avec le modèle de la section précédente dans le chapitre suivant, le modèle retenu sera celui sélectionné sur la base du critère *BIC* également.

Les différentes validations effectuées sont présentées dans les prochains paragraphes.

En premier lieu, l'analyse du graphique des résidus en fonction des valeurs prédites suggère que les résidus sont distribués aléatoirement autour de zéro. De plus, le graphique présentant les résidus en fonction des quantiles de la loi normale montre une ligne droite sur une importante partie de la courbe, à l'exception d'une dizaine de points qui contribuent à faire courber l'extrémité gauche et une quinzaine de points qui contribuent à faire courber l'extrémité droite. Par rapport au nombre total d'observations, ce nombre de points est très petit, ce qui porte à conclure qu'il est plausible que la distribution des erreurs soit suffisamment près d'une loi normale pour se fier aux valeurs des paramètres estimés et aux seuils observés obtenus par le modèle. Les graphiques analysés sont présentés en seconde partie de l'Annexe 8.

Ensuite, l'observation des graphiques de résidus en fonction des variables dépendantes permet de détecter une présence potentielle d'hétéroscédasticité pour cinq variables, soient :

- le montant total de ventes nettes virtuelles (1);
- le montant médian autorisé pour l'ensemble des prêts (2);
- la proportion de clients détenant un produit d'épargne 3 (3);
- la proportion de clients habitant à l'extérieur d'une région métropolitaine, aucune influence métropolitaine (4);
- l'écart-type du montant détenu dans le produit d'épargne 4 (5).

Effectivement, la distribution des résidus en forme d'entonnoir sur les graphiques semble suggérer que ce type de problème est présent. De plus, le seuil observé plus faible que 0,0001 du test obtenu avec l'option *SPEC* de *PROC REG* entraine le rejet de l'hypothèse nulle d'homogénéité des variances au seuil de 5 %. Afin de remédier à ce problème, des transformations sur ces variables sont tentées, ainsi que des transformations sur la variable dépendante. La transformation « racine carrée » appliquée sur quatre variables explicatives (1 à 4) est celle qui donne les meilleurs résultats pour ces variables et la transformation log semble donner un meilleur résultat pour la cinquième variable (5), tant au niveau de l'amélioration de la distribution des résidus constatée par l'observation des graphiques que sur le résultat du test de l'homogénéité des variances dont le seuil observé passe à 0,3195. Finalement, un diagnostic de la multicolinéarité avec l'utilisation de l'option *COLLINOINT* de la procédure *REG* est opéré. Comme les valeurs de l'indicateur « *Condition Index* » du tableau « *Collinearity Diagnostics (intercept adjusted)* » sont toutes inférieures à 3, aucun problème de multicolinéarité ne semble présent dans ce modèle. De plus, toutes les valeurs de l'indicateur de l'inflation de la variance associées aux variables explicatives sont inférieures à 2, ce qui confirme l'absence d'un problème de multicolinéarité. En guise de complément de validation portant sur la détection d'un problème de multicolinéarité, des modèles univariés avec chacun des effets fixes sélectionnés sont comparés avec le modèle complet, afin de

valider si le signe des paramètres estimés est le même dans les deux cas. Étant donné que tous les signes des paramètres estimés demeurent les mêmes, les résultats obtenus sont confirmés.

Au terme de ces validations et des corrections appropriées, la valeur du critère *BIC* diminue (-13114,1 contre -13108,6 précédemment), ce qui permet de constater que les transformations de variables permettent d'améliorer le modèle.

Dans la section traitant de l'intégration des données de site web, il était mentionné qu'une migration d'outil avait eu lieu en septembre 2013, ce qui pouvait avoir un impact important sur la façon de calculer les indicateurs et les filtres utilisés. Ainsi, la création d'une variable indicatrice permettant de distinguer les trimestres avant la migration et après la migration est créée. Cette variable permet de vérifier si l'ajout d'une interaction entre cette variable indicatrice et la variable de type web présente dans le modèle permet d'améliorer le modèle selon la valeur du critère *BIC*. L'évaluation de l'intégration de cette variable indicatrice ne fait pas partie du processus de sélection, puisqu'il s'agit d'une variable temporaire, qui, à terme, ne sera pas utilisée. En effet, si le modèle est ajusté dans le futur, la contrainte de migration d'outil ne sera pas d'actualité. En ajustant un nouveau modèle où l'interaction entre la variable indicatrice et la variable de type web remplace la variable de type web, la valeur du critère *BIC* demeure pratiquement la même (augmente légèrement), ce qui permet de déduire que le modèle n'est pas affecté par l'ajout de cette information. Ainsi, ces résultats portent à croire que les changements apportés en septembre 2013 auraient causé beaucoup moins d'impacts que ce qui était attendu, du moins en ce qui a trait aux variables utilisées dans le modèle et aux filtres appliqués sur les données de l'outil *Google Analytics*. En conclusion, aucun changement n'est apporté au modèle relativement aux données de type web.

Intégration de certains termes d'interaction et d'ordre supérieur

Comme dans le cas du modèle construit avec les données des entrepôts internes, il est possible de préciser le modèle obtenu au terme de l'étape des validations. Les détails concernant la sélection des effets fixes d'ordre supérieur sont présentés à l'Annexe 7.

Le processus d'ajout d'effets fixes d'ordre supérieur a permis de sélectionner six effets fixes supplémentaires représentant des interactions entre deux variables. En effet, la valeur du *BIC* a constamment diminué jusqu'à ce nombre, pour ensuite remonter. L'ajout de ce grand nombre d'effets fixes permet d'améliorer légèrement le niveau de précision du modèle, la valeur du critère *BIC* selon la méthode *ML* passant de -13114,1 à -13127,6. En considérant un objectif de prévision, l'ajout d'interactions et de termes d'ordre supérieur constituerait une certaine amélioration.

Pour terminer, la validation de la structure de covariance sélectionnée au départ sur le modèle complet a permis de constater qu'elle est toujours celle qui donne les meilleurs résultats en ce qui concerne les valeurs obtenues au niveau des critères de sélection.

5.2 Comparaison des modèles 1 et 2

L'objectif principal de ce projet de recherche étant de déterminer si l'ajout de données de sources variées non habituellement utilisées au sein d'une entreprise du secteur bancaire aux données des entrepôts internes permet d'améliorer la performance de modèles statistiques, il importe de comparer les modèles sélectionnés dans la section précédente. Cette comparaison peut se faire selon les critères de sélection *AIC* et *BIC*, où la plus faible valeur témoigne d'un modèle plus précis, mais également sur la base des types de variables sélectionnées, dans un contexte d'interprétation de modèles explicatifs.

En premier lieu, les critères *AIC* et *BIC* permettent de constater que les modèles sélectionnés dont les données proviennent de sources multiples de données (modèle 2) sont plus performants que les modèles sélectionnés dont les données proviennent uniquement des entrepôts internes (modèle 1), sauf lorsque les termes d'interactions sont inclus dans le modèle. Ces observations sont valides à la suite du processus de sélection d'effets fixes suivant la présélection de variables et après les validations et corrections à apporter pour régler certains problèmes. Ainsi, si le modèle est utilisé à des fins explicatives, le modèle construit à l'aide de variables provenant de sources multiples de données est légèrement meilleur du point de vue des critères de sélection. Toutefois, en fonction des résultats obtenus dans le cas d'application, un développement de modèles à des fins de prédiction aurait avantage à utiliser des variables des entrepôts de données internes, étant donné l'amélioration significative que procure l'ajout d'interactions et de termes d'ordre supérieur du point de vue du critère de sélection *BIC*. Le Tableau 9 permet de constater ces résultats.

Tableau 9 : Comparaison des modèles sélectionnés au cours du processus de sélection, pour les deux types de modèles

	Meilleur modèle selon le critère	
	AIC	**BIC**
Sélection des effets fixes		
Modèle 1	-13274,1	-13105,1
Modèle 2	**-13298,3**	**-13108,6**
Validations et corrections		
Modèle 1	s. o.	-13110,1
Modèle 2	s. o.	**-13114,1**
Ajout d'interactions et de variables au carré		
Modèle 1	s. o.	**-13198,8**
Modèle 2	s. o.	-13127,6

Alors que les critères de sélection permettent une comparaison de modèles au niveau de leur performance statistique, la comparaison des différents types de données incluses dans les modèles est importante dans une perspective de compréhension du contexte d'affaires. En effet, même si un modèle est performant du point de vue statistique, il peut être en mesure de donner peu d'informations concrètes relatives à la situation réelle de l'entreprise dans le contexte analysé.

Le modèle 1 (données des entrepôts internes) sélectionné sur la base du critère *BIC* met en relation des variables de quatre catégories pour l'explication de la performance des ventes nettes en épargne dans les succursales, soient temporelle, financière, sociodémographique et comportementale. Malgré le plus grand nombre de variables considérées dans le modèle sélectionné sur la base du critère *AIC,* des variables des mêmes catégories sont représentées.

En ce qui a trait au modèle 2 (données de multiples sources), les variables considérées dans le modèle sélectionné sur la base du critère *BIC* proviennent de plusieurs catégories,

soient temporelle, financière, sociodémographique, géographique et web. En considérant le modèle sélectionné sur la base du critère AIC, la catégorie sollicitation s'ajoute. Ce vaste éventail de variables composant les modèles permet de faire une interprétation plus riche des éléments qui sont déterminants dans l'analyse de la performance de succursales en lien avec les ventes en épargne. Le modèle 2 sélectionné sur la base du critère *BIC* comprend un moins grand nombre de variables, ce qui facilite l'interprétation. De plus, il n'est pas associé à un problème de multicolinéarité, ce qui présente un autre avantage du point de vue de l'interprétation.

Plusieurs sources de données étaient considérées dans la construction du modèle 2. Rappelons que ces sources étaient les suivantes :

- entrepôts de données internes;
- Statistique Canada;
- organismes de santé publique;
- prestation de service;
- sollicitation et marketing direct;
- site web.

Des variables provenant de quatre des six sources de données ont été sélectionnées dans le modèle 2, en considérant le critère *AIC*. Les sources n'étant pas représentées sont l'outil de prestation de service et les organismes de santé publique. Le modèle sélectionné sur la base du critère BIC n'intégrait pas de variables de sollicitation et marketing direct. Ces variables, tel qu'elles ont été construites, ne permettaient pas un apport significatif au modèle permettant d'expliquer la performance des succursales sur le plan des ventes en épargne telle qu'elle est mesurée. Même lors de l'étape de présélection de variables, les variables provenant des organismes de santé publique n'ont pas été sélectionnées. Le Tableau 10 résume le nombre de variables sélectionnées par source, selon les meilleurs modèles sélectionnés sur la base des critères *AIC* et *BIC*.

Tableau 10 : Nombre de variables sélectionnées par source, selon les critères *AIC* et *BIC*

	Nombre de variables sélectionnées	
Source de données	**Critère AIC**	**Critère BIC**
Entrepôts de données internes	24	8
Statistique Canada	2	1
Organismes de santé publique	0	0
Prestation de service	0	0
Sollicitation et marketing direct	1	0
Site web	3	2

5.3 Validation de la méthode de sélection de modèles utilisée

Dans le cadre du processus de sélection de variables et de modèles, une méthodologie a été proposée et développée. Cette méthodologie visait à tenir compte du contexte de mesures répétées relié au cas d'application tout au long du processus de sélection. De plus, cette méthode se voulait être un juste milieu entre une méthode de sélection par la comparaison de tous les modèles possibles, impossible en tenant compte des mesures répétées avec les ressources informatiques disponibles, et une méthode de type *STEPWISE*. Cette méthode a été utilisée pour le choix de deux modèles quant au cas d'application, soient un modèle utilisant des données des entrepôts de données internes de l'entreprise et un modèle utilisant des données provenant de sources multiples. Afin de valider si cette méthode donne de bons résultats et permet de sélectionner de bons modèles, une méthode de sélection de modèles automatisée disponible dans *SAS* a été utilisée. Ainsi, l'utilisation de la procédure *REG* a permis de comparer tous les modèles possibles pour tous les nombres de variables dans le modèle à partir des variables présélectionnées (sélection de type *all-subset*). Plus précisément, la procédure permet de sélectionner le meilleur modèle pour chaque nombre de variables possibles dans le modèle sur la base de la statistique R^2, tout en retenant les valeurs associées aux critères de sélection *AIC, BIC* et *SBC (Swartz Bayesian Information Criterion)*. Par la suite, les plus faibles valeurs obtenues pour ces critères parmi tous les modèles confondus permettent de sélectionner le meilleur modèle selon le critère choisi. L'ajustement des

modèles choisis selon ces critères dans la procédure *MIXED* afin de tenir compte des mesures répétées permet d'obtenir les valeurs associées aux critères *AIC* et *BIC* pouvant être comparées à celles obtenues au terme du processus de sélection de modèles utilisé dans le cas d'application. Le Tableau 11 permet de comparer les résultats obtenus à l'aide des deux méthodes, selon les critères de sélection *AIC* et *BIC*. L'observation de ces résultats permet de constater que la méthode proposée et utilisée dans le cas d'application a permis de sélectionner en général de meilleurs modèles que la méthode de sélection automatisée en évaluant tous les modèles possibles, et ce, peu importe le critère de sélection considéré. Dans le cas d'une sélection sur la base du critère *BIC*, les avantages sont plus marqués que lorsque le critère *AIC* est utilisé. En effet, du point de vue du critère *AIC*, pour le modèle 1, la méthode proposée est associée au meilleur modèle, alors que pour le modèle 2, les résultats sont très similaires au modèle sélectionné de façon automatisé en considérant le critère *AIC*.

Tableau 11 : Comparaison de la méthode de sélection utilisée avec une méthode de sélection automatisée de SAS

		Proc mixed (ml)	
	Nombre de variables	**AIC**	**BIC**
Modèle 1 : Données provenant des entrepôts internes de l'entreprise			
Sélection « all-subset » – critère AIC	34	-13257,6	-13012,8
Sélection « all-subset » – critère BIC	31	-13254,4	-13021,2
Sélection « all-subset » – critère SBC	5	-13160,3	-13028,2
Sélection proposée – critère AIC	29	**-13274,1**	
Sélection proposée – critère BIC	12		**-13105,1**
Modèle 2 : Données provenant de sources multiples			
Sélection « all-subset » – critère AIC ou BIC	31	**-13279,1**	-13045,9
Sélection « all-subset » – critère SBC	9	-13152,1	-13004,4
Sélection proposée – critère AIC	32	-13278,3	
Sélection proposée – critère BIC	26		**-13108,6**

5.4 Interprétation d'un modèle sélectionné

À l'issue des étapes de sélection de modèles et de comparaison entre les modèles 1 et 2, il appert que le modèle 2, où des données de sources multiples sont utilisées, performe mieux du point de vue des critères *AIC* et *BIC*, sans tenir compte des interactions et des effets d'ordre supérieur. De plus, la variété des types de variables utilisés dans le modèle rend son interprétation intéressante. La version du modèle 2 sélectionné à l'aide du critère *BIC* est utilisée pour interprétation.

Tout d'abord, les coefficients estimés obtenus en ajustant le modèle permettent de comprendre les effets positifs et négatifs des variables sur la performance des succursales au niveau des ventes en épargne. Ces coefficients obtenus pour les variables explicatives sont présentés dans le Tableau 12 et sont classés en fonction du type d'effet sur la variable réponse, soit positif ou négatif. Dans ce tableau, les seuils observés permettent de constater que les coefficients estimés sont tous significatifs au seuil de 2%.

Tableau 12 : Classification des variables en fonction de leur type d'effet sur la variable réponse, modèle final 2 sans interactions

Variables	Coefficients estimés	Erreurs standard	Seuils observés	Sources de données
Effet positif sur la variable réponse				
Durée moyenne des sessions dans le groupe d'heures 2	0,000008	0,000002	0,0002	Google Analytics
Montant total de ventes nettes virtuelles (racine carrée)	0,000004	0,000000	<0,0001	Entrepôts internes
Proportion de clients dont l'état civil est « célibataire »	0,070009	0,014894	<0,0001	Entrepôts internes
Montant médian autorisé pour l'ensemble des prêts (racine carrée)	0,000550	0,000227	0,0161	Entrepôts internes
Proportion de clients dont l'état civil est « initial »	0,058950	0,013706	<0,0001	Entrepôts internes
Proportion de clients appartenant à une communauté culturelle	0,025346	0,006422	<0,0001	Entrepôts internes
Proportion de clients détenant un produit d'épargne 3 (racine carrée)	0,047571	0,012273	0,0001	Entrepôts internes
Effet négatif sur la variable réponse				
Temps	-0,006280	0,000257	<0,0001	Entrepôts internes
Ratio nombre de sessions dans le groupe d'heures 3 / nombre de sessions totales	-0,016058	0,006643	0,0161	Google Analytics
Proportion de clients habitant à l'extérieur d'une région métropolitaine, aucune influence métropolitaine (racine carrée)	-0,022746	0,006312	0,0004	Statistique Canada
Écart-type du montant détenu dans le produit d'épargne 4 (log)	-0,006313	0,001443	<0,0001	Entrepôts internes

Variables ayant un effet positif

Sept variables ont un effet positif sur la performance de la succursale. Ces variables sont de nature financière, sociodémographique et web.

Du point de vue financier, une hausse du montant moyen par client de ventes nettes en produits virtuels, une augmentation du montant médian autorisé pour l'ensemble des prêts ou une augmentation de la proportion de clients détenteurs d'un produit d'épargne 3 sont des éléments qui ont un effet positif sur le ratio de ventes nettes par rapport au volume d'affaires total de la succursale.

L'effet positif des produits de crédit sur les ventes nettes en épargne est intéressant. En effet, on pourrait penser que la détention de produits de crédit est inversement proportionnelle à la détention de produits d'épargne. Or, les individus dont le montant autorisé en produits de crédit est élevé ont probablement tendance à avoir un revenu plus élevé. Lorsque le montant moyen autorisé en produits de crédit augmente pour une succursale, on peut supposer que les ressources financières des clients augmentent, ce qui conduit à un meilleur potentiel au niveau de l'épargne. De plus, le fait de faire une demande de crédit témoigne d'une relation d'affaires plus importante avec l'entreprise. Lorsque le montant médian autorisé pour l'ensemble des prêts augmente de 10 000$ tout en gardant les autres variables fixes, le ratio de ventes nettes par rapport au volume d'affaires de la succursale augmente en moyenne de 0,06.

Au niveau de la vente de produits virtuels, l'effet positif se traduit par une augmentation moyenne du ratio à expliquer de 0,004 lorsque ces ventes augmentent de 1 million de dollars dans un trimestre pour une succursale, si toutes les autres variables explicatives sont fixes. Cet effet positif est probablement lié au fait que des offres fortement concurrentielles ont été développées dans les dernières années, ce qui a permis aux

succursales dont les clients ont des habitudes d'investissement en ligne d'augmenter leur nombre de clients de produits d'épargne.

Le produit d'épargne 3 est un produit d'épargne fluctuante, très commun dans le portefeuille des clients. Ainsi, une proportion élevée de ce produit peut être interprétée comme une proportion élevée d'épargnants dans la succursale. Selon le modèle étudié, une augmentation de 0,01 de la proportion de clients détenteurs du produit d'épargne 3 aurait un impact moyen de 0,005 sur le ratio de ventes nettes par rapport au volume d'affaires de la succursale, lorsque toutes les autres variables sont considérées comme fixes.

Du point de vue sociodémographique, deux éléments ont un effet positif sur le ratio de ventes nettes par rapport au volume d'affaires de la succursale, soient l'état civil et l'appartenance à une communauté culturelle.

En ce qui a trait à l'état civil, une augmentation de la proportion de clients célibataires, lorsque toutes les autres variables sont considérées comme fixes, a un impact positif sur le ratio. Plus précisément, une augmentation de 10% du nombre de clients célibataires par rapport au nombre total de clients a un impact moyen positif de 0,007 sur le ratio de ventes par rapport au volume d'affaires de la succursale, lorsque toutes les autres variables indépendantes sont constantes. Un effet similaire, mais légèrement moins prononcé, s'applique en considérant la proportion de clients dont l'état civil n'est pas connu (initial).

Selon le modèle, une proportion plus élevée de clients appartenant à une communauté culturelle a un effet positif sur le ratio témoignant de la performance de la succursale en regard de l'épargne. En effet, une proportion 10% plus élevée du nombre de clients appartenant à une communauté culturelle a un effet moyen de 0,003 sur le ratio de ventes

nettes sur le volume d'affaires de la succursale, lorsque toutes les autres variables sont stables. Certains éléments pourraient permettre d'expliquer un tel effet : certaines cultures plus propices à l'épargne, fidélité à l'entreprise plus marquée pour certaines communautés culturelles, arrivée récente au pays de certains clients ayant pour conséquence que des achats immobiliers ne se sont pas concrétisés.

Une variable de type web a un effet positif sur le ratio de ventes nettes sur le volume d'affaires de la succursale, soit la durée moyenne des sessions sur le site web de l'entreprise entre 7h et 16h. En effet, lorsque cette durée moyenne augmente, mais que toutes les autres variables restent stables, le ratio de performance augmente. Plus précisément, une augmentation de 1,7 minute (100 secondes) de la durée moyenne alors que toutes les autres variables demeurent stables aurait un impact moyen de 0,0008 sur le ratio de performance. Différentes hypothèses peuvent être posées pour expliquer cet effet : les visites sur le site web de l'entreprise pendant le jour sont faites par des clients ayant un objectif sérieux de nature financière, les visites de jour témoignent d'une plus grande accessibilité à Internet ou au site web de l'entreprise pour les clients, les clients visitant le site de jour bénéficient de plus de temps.

Variables ayant un effet négatif

Quatre variables, toutes de catégories différentes, ont un effet négatif sur le ratio à expliquer. Ces variables sont de nature temporelle, web, géographique et financière.

Tout d'abord, l'évolution dans le temps a un effet négatif sur le ratio de ventes nettes par rapport au volume d'affaires de la succursale, lorsque toutes les autres variables explicatives sont considérées comme fixes. En effet, une évolution d'un trimestre a un impact moyen négatif de 0,006 sur le ratio à expliquer, lorsque toutes les autres variables demeurent constantes. Par contre, une croissance rapide du volume d'affaires de la

succursale en raison du développement relié à d'autres produits financiers pourrait expliquer la tendance légèrement à la baisse de ratio à expliquer.

Au niveau du comportement sur le site web de l'entreprise, le coefficient estimé pour le ratio du nombre de sessions en soirée par rapport au nombre total de sessions suggère qu'une augmentation du poids des fréquentations en soirée a un effet négatif sur le ratio de ventes nettes par rapport au volume d'affaires de la succursale. Plus précisément, une augmentation de la proportion des fréquentations en soirée de 10% a un impact moyen négatif de 0,002 sur le ratio expliqué, lorsque toutes les autres variables demeurent constantes. Certaines hypothèses pourraient permettre d'expliquer cette influence : les visites de soirée correspondent à des clients potentiels moins sérieux, les visites d'individus en recherche de concours ou de promotions se font surtout de soir, la fréquentation de soirée pourrait être plus de nature informative ou éducative.

Le coefficient estimé correspondant à la proportion de clients habitant à l'extérieur d'une région métropolitaine et n'ayant aucune influence métropolitaine, une variable de type géographique, suggère qu'un impact négatif est causé par cette variable sur le ratio de performance. En effet, en considérant toutes les autres variables fixes, une augmentation de 1% de la proportion de clients habitant dans ce type de région a un effet négatif moyen de 0,005 sur la variable dépendante. Quelques hypothèses peuvent aider à expliquer cet effet : précarité d'emploi, éducation financière moins marquée, ressources financières limitées, personnel en succursale moins spécialisé, manque d'accessibilité à des informations financières.

L'écart-type du montant détenu en produit d'épargne 4 permet de constater de la variabilité au niveau de cet indicateur dans chaque succursale. Ainsi, un écart-type plus grand témoigne d'une plus grande variation pour les différents montants détenus par rapport à la moyenne. Lorsque la variabilité du montant détenu dans le produit d'épargne augmente, mais que les autres variables demeurent stables, l'effet sur le ratio de ventes

nettes par rapport au volume d'affaires de la succursale est négatif. Des enjeux de rétention relativement à ce produit d'épargne sont vécus dans l'entreprise depuis quelques années, ce qui peut permettre d'expliquer ce comportement dans le modèle. En effet, certaines succursales pourraient avoir plus de difficultés que d'autres à limiter les sorties de fonds, ce qui implique une conséquence directe négative sur les ventes nettes.

Chapitre 6 : Discussion

L'analyse des résultats a démontré que l'utilisation de sources multiples de données pour le développement de modèles statistiques permettant d'expliquer la performance des ventes en épargne dans des succursales d'une entreprise du secteur bancaire permet de construire des modèles légèrement plus performants, autant en considérant des critères de sélection statistique que les possibilités d'interprétation des modèles. Toutefois, il importe de considérer plusieurs éléments avant de conclure que cette pratique doit être valorisée dans tous les cas de développement de modèles de ce type. La présente section permet de mettre en lumière des éléments qui doivent essentiellement être soulignés.

L'utilisation de données provenant de multiples sources requiert la considération d'une démarche d'intégration de données, qu'elle soit récurrente et automatisée ou non. Cette démarche d'intégration nécessite d'importants investissements de natures diverses : ressources humaines, informatiques et logicielles, temps. Ces investissements doivent être estimés et quantifiés avant de prendre la décision d'utiliser une telle démarche pour le développement de modèles. Effectivement, les gains prévus relatifs à la performance des modèles doivent compenser les efforts requis. Cette préoccupation est d'autant plus cruciale dans le cas d'un processus non développé et non structuré à priori. Afin de prendre une décision éclairée, il importe également de se questionner sur la récurrence et la mise en production éventuelle du modèle à développer. Dans ce cas, la mise en production d'un traitement d'intégration des données requises est sans aucun doute essentielle. En effet, un processus structuré d'intégration de données permettrait de diminuer l'impact de certaines contraintes vécues au cours de ce projet de recherche : difficulté d'obtenir les données en temps opportun, secteurs de l'entreprise moins collaborateurs, sources de données de formats variables et non constants, identifiants

différents en fonction des sources de données, compréhension exigeante et nécessaire de certaines sources de données.

Dans ce projet de recherche, l'intégration de données qualitatives portant sur les activités de marketing direct et de sollicitation a été tentée afin de vérifier si ces informations pouvaient permettre d'expliquer la performance des succursales en regard des ventes en épargne à l'aide de modèles statistiques. Or, toutes les variables construites à partir de ces données n'ont pas été sélectionnées dans les modèles, ce qui indique que leur potentiel explicatif était limité par rapport aux autres variables utilisées et à la variable réponse. Toutefois, avant d'exclure ce type de variables d'un projet futur similaire, il importe de se questionner sur la façon dont les variables ont été construites. En effet, peut-être que le développement d'indicateurs différents permettrait d'obtenir de meilleurs résultats du point de vue de ce type de données? Cette tentative d'intégration de ce type de données constituait une nouveauté par rapport à l'exploitation de ces données au sein de l'entreprise et mériterait d'être renouvelée sous un angle différent. Les efforts dans ce sens pourraient être étendus également vers l'intégration de données reliées à des activités publicitaires et promotionnelles.

Les données provenant de l'outil de prestation de service sont très prisées au sein de l'entreprise. En effet, leur accessibilité est récente et les données sont encore en cours d'exploration. Les indicateurs utilisés dans ce cas d'application sont ceux qui étaient développés au moment où les données étaient requises. Malheureusement, aucun de ces indicateurs n'a eu suffisamment de poids pour expliquer la variabilité dans les modèles sélectionnés. Or, quelques analyses au sein de l'entreprise tendent à démontrer que l'utilisation adéquate de l'outil de prestation de service permet d'améliorer les résultats de ventes en épargne. Il serait donc fort approprié de poursuivre les tentatives d'intégration de ce type de données dans des modèles statistiques concernant la performance des succursales au niveau des ventes en épargne. Les indicateurs et variables utilisés peuvent sans aucun doute être raffinés.

Les résultats de l'intégration des données provenant de l'outil *Google Analytics* sont positifs. En effet, quelques variables ont permis d'apporter de nouvelles informations pertinentes pour expliquer la performance des succursales. L'intégration de ce type de données à l'aide d'éléments géographiques constituait une première par rapport à leur utilisation au sein de l'entreprise. Ces données sont également associées aux contraintes et difficultés les plus importantes vécues dans ce projet de recherche. Ce succès d'utilisation de données provenant de la fréquentation du site web de l'entreprise porte à croire qu'il y a un important potentiel d'exploitation relativement à ces données. Ainsi, des travaux à partir des données non agrégées sont fortement recommandés, afin d'aller plus loin au niveau de l'appariement des données et des variables pouvant être utilisées. L'accessibilité à ce type de données étant très limitée au sein de l'entreprise, une intégration aux entrepôts de données devrait être envisagée, étant donné le potentiel d'exploitation pressenti.

Les sources utilisées dans le cas d'application constituent un pas dans la direction de l'analyse et l'exploitation de données volumineuses (*Big data*) pour l'entreprise. L'intégration d'autres sources de données telles que des données provenant des médias sociaux, des agendas de rendez-vous et des médias financiers permettraient de poursuivre les développements initiés.

La méthode de sélection de modèles et de variables développée donne de bons résultats dans le contexte du cas d'application étudié. Il serait intéressant de vérifier si cette méthode peut être associée à des résultats similaires dans des contextes différents, pour éventuellement être généralisée. En contrepartie, en considérant la façon dont elle est automatisée, cette méthode requiert beaucoup de ressources informatiques, d'intervention humaine et de temps lorsque le nombre de variables est grand. Il serait donc pertinent de faire évoluer le traitement SAS pour le rendre plus simple d'utilisation et moins contraignant sur le plan des ressources requises, notamment.

Les modèles développés dans le contexte de cette recherche visaient principalement l'explication de la performance des succursales en regard des ventes en épargne. En contrepartie, une ébauche pour la construction d'un modèle prédictif a été faite. En effet, l'ajout d'interactions et de termes d'ordre supérieur permet d'améliorer la précision des modèles de façon non négligeable, ce qui constituerait un ajout positif en matière de prédiction. Étant donné que toutes les variables explicatives ne sont pas disponibles un trimestre à l'avance, le développement d'un modèle prédictif devrait se faire en décalant les variables explicatives d'un trimestre par rapport à la variable réponse. De plus, la présence de multicolinéarité ne représentant pas un problème dans le cas d'un modèle prédictif, le nombre de validations à effectuer sur le modèle serait réduit. Finalement, il serait pertinent de considérer des modèles plus complexes (comme ceux sélectionnés sur la base du critère *AIC*) et de les comparer avec des modèles plus parcimonieux à l'aide de méthodes de validation, telle que la validation croisée.

Le développement de modèles statistiques peut porter sur de multiples usages. En effet, outre le volet de compréhension de la performance, les modèles peuvent être utilisés pour la fixation d'objectifs de ventes, pour prédire la performance future des succursales, pour cibler les succursales en difficulté et ainsi leur apporter plus de soutien pour la réalisation de leurs activités d'affaires et pour orienter le développement de produits adaptés à leur réalité. En fonction de l'objectif principal considéré lors du développement d'un modèle, il peut être nécessaire de sélectionner un modèle plus précis, mais plus complexe, ou un modèle permettant une interprétation aisé, mais moins performant. Le juste milieu entre performance statistique et contexte d'affaires doit être recherché.

Conclusion

La présente recherche a permis de démontrer que l'utilisation de données provenant de multiples sources permet d'améliorer la performance des modèles statistiques, dans un contexte d'analyse de la performance des ventes en produits d'épargne dans des succursales d'une entreprise du secteur bancaire. Afin de répondre à la question de recherche, plusieurs éléments ont été considérés. Tout d'abord, une réflexion sur les sources de données à utiliser a été effectuée. Ensuite, les concepts d'intégration de données et d'appariement ont été largement explorés. Par la suite, les modèles statistiques utilisés dans un contexte de mesures répétées, ainsi que les méthodes les plus adéquates pour procéder à la sélection de modèles et de variables, lorsque le nombre de variables potentielles est élevé et la taille de l'échantillon est grande, sont des éléments qui ont été travaillés. D'ailleurs, par rapport à ce dernier point, le développement d'une méthode de sélection de variables permettant de tenir compte du contexte de mesures répétées a permis d'effectuer une sélection de modèles plus performants sur la base de critères de sélection qu'une méthode automatisée, habituellement utilisée dans ce type de contextes, accessible dans SAS. Finalement, plusieurs éléments de développements ont été discutés quant à l'utilisation de données provenant de sources multiples et certaines sources ont été jugées prometteuses pour des analyses futures.

Annexes

Annexe 1 : Éléments SAS développés pour la sélection de variables ou de modèles dans un contexte de mesures répétées

```
libname temp_mod "ma_librairie";

%let variables_class = mes variables de classification;

*** Sélection parmi tous les modèles possibles;
%macro possibles(toutesvar);
proc mixed data = memoirecomplet method=ml ic;
        class &variables_class.;
        model y = &toutesvar. / solution cl;
        repeated temps_cat / subject=succursale type=un;
        ods output InfoCrit = criteres;
run;

data criteres;
        set criteres;
        toutesvar = "&toutesvar.";
run;

proc append data=criteres base=temp_mod.totinfocrit;
run;
%mend possibles;

*****************************************************************************************;
*** Modèles à 1 effet fixe - modèle 2;
*****************************************************************************************;

data temp_mod.totinfocrit;
        length toutesvar $60. Neg2LogLike parms aic aicc hqic bic caic 8.;
        toutesvar = ' ';
        Neg2LogLike = .;
        parms = .;
        aic = .;
        aicc = .;
        hqic = .;
        bic = .;
        caic = .;
run;

ods select none;
data _null_;
        set var_preselect_2;
        call execute ('%possibles(' || nomvar || ')');
run;
ods select all;

%macro meilleur(critere,nomdata);
proc sort data = temp_mod.totinfocrit (where = (toutesvar ne " "));
        by &critere.;
run;

data &nomdata.;
        set temp_mod.totinfocrit (obs = 1);
run;
%mend meilleur;

%meilleur(aic, temp_mod.meilleurAIC2_1effetsfixes);


*****************************************************************************************;
*** Modèles à 2 effets fixes - modèle 2;
*****************************************************************************************;
```

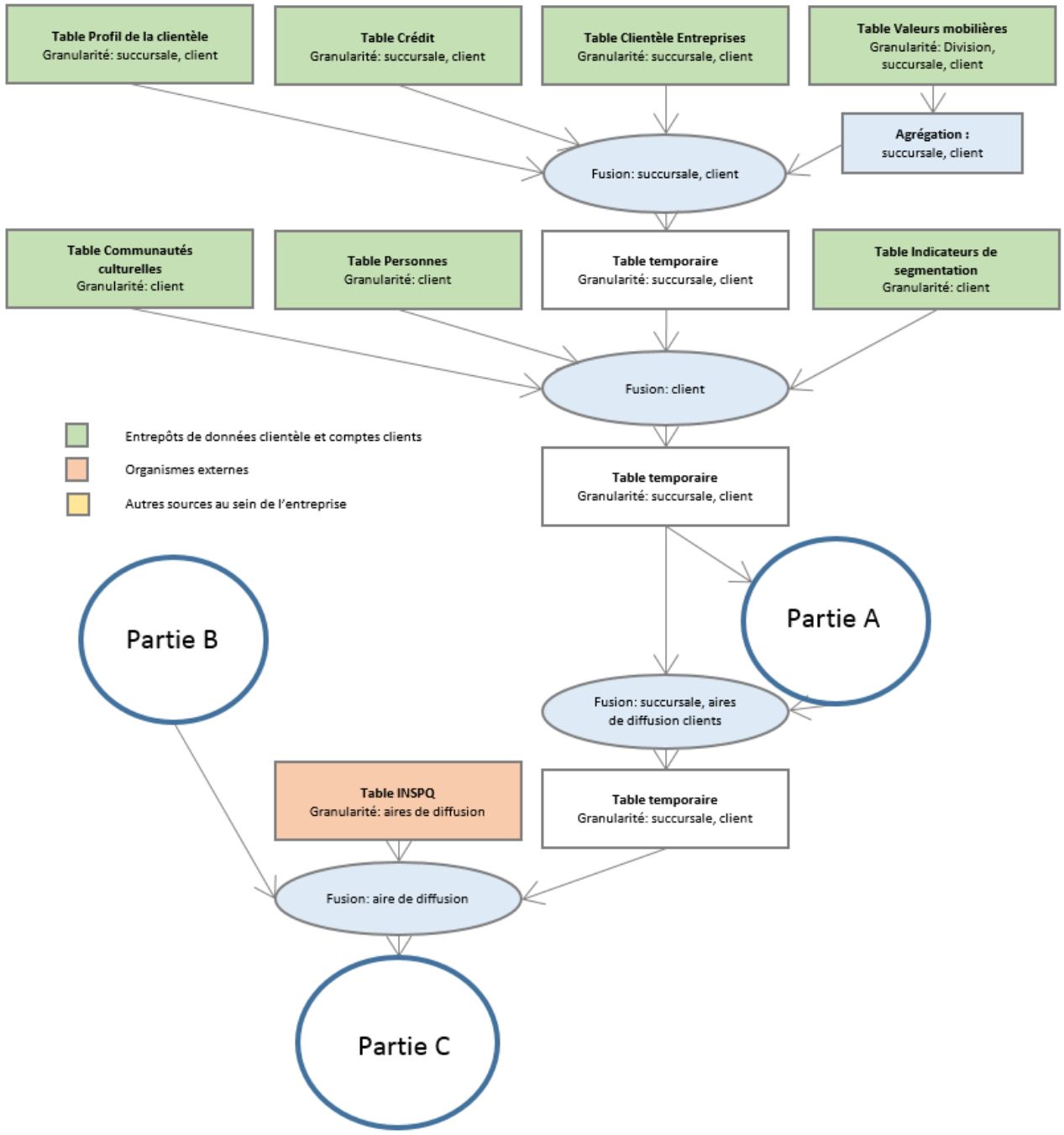
```
proc sql;
	create table temp_mod.select2_2 (where = (nomvar1 ne nomvar2)) as
		select *
		from var_preselect_2 (rename = (nomvar = nomvar1)),
		     var_preselect_2 (rename = (nomvar = nomvar2))
	;
quit;

data temp_mod.select2_2a;
	set temp_mod.select2_2;
	if nomvar1 < nomvar2 then toutesvar = compress(nomvar2) || " " ||
compress(nomvar1);
	else if nomvar2 <= nomvar1 then toutesvar = compress(nomvar1) || " " ||
compress(nomvar2);
run;

proc sort data = temp_mod.select2_2a out = temp_mod.select2_2a nodupkey; by toutesvar;
run;

data temp_mod.totinfocrit;
	length toutesvar $60. Neg2LogLike parms aic aicc hqic bic caic 8.;
	toutesvar = ' ';
	Neg2LogLike = .;
	parms = .;
	aic = .;
	aicc = .;
	hqic = .;
	bic = .;
	caic = .;
run;

ods select none;
data _null_;
	set temp_mod.select2_2a;
	call execute ('% possibles (' || toutesvar || ')');
run;
ods select all;

%meilleur(aic, temp_mod.meilleurAIC2_2effetsfixes);

/*
Description : Permets de constituer un fichier donnant toutes combinaisons de modèles
possibles,
	selon la méthodologie développée. La complexité de ce traitement est causée par
le retrait des
	doublons de variables, peu importe leur ordre. Le retrait des doublons permet de
tester un
	nombre raisonnable de modèles.
	Peut être utilisée pour des modèles à partir de trois effets fixes.


no_modele = 1 ou 2
fixe=1 si variable fixée pour l'étape. 0 sinon.
prec2=1 si deux variables ont été sélectionnées à l'étape précédente. 0 si 3 variables
ont été sélectionnées.
nb_var=nombre de variables considérées pour le choix du meilleur modèle
ajout_fixe1=" nouvelle variable fixée 1 (s'il y a lieu)    "
ajout_fixe2=" nouvelle variable fixée 2 (s'il y a lieu) "
*/


%macro allposs2(no_modele, fixe, prec2, nb_var, ajout_fixe1, ajout_fixe2);

%if &fixe.=1 and &prec2.=1 %then %do;
	data temp_mod.select&no_modele._%eval(&nb_var.-1)a (drop = nomvar%eval(&nb_var.-2)0
nomvar%eval(&nb_var.-1)0);
		set temp_mod.select&no_modele._%eval(&nb_var.-1)a (rename = (nomvar%eval(&nb_var.-
2) = nomvar%eval(&nb_var.-2)0
```

```
                        nomvar%eval(&nb_var.-1) = nomvar%eval(&nb_var.-1)0));
                if nomvar%eval(&nb_var.-2)0 = &ajout_fixe1. then do;
                        nomvar%eval(&nb_var.-2) = nomvar%eval(&nb_var.-2)0;
                        nomvar%eval(&nb_var.-1) = nomvar%eval(&nb_var.-1)0;
                end;
                else if nomvar%eval(&nb_var.-1)0 = &ajout_fixe1. then do;
                        nomvar%eval(&nb_var.-2) = nomvar%eval(&nb_var.-1)0;
                        nomvar%eval(&nb_var.-1) = nomvar%eval(&nb_var.-2)0;
                end;
                else do;
                        nomvar%eval(&nb_var.-2) = nomvar%eval(&nb_var.-2)0;
                        nomvar%eval(&nb_var.-1) = nomvar%eval(&nb_var.-1)0;
                end;
        run;
%end;

%else %if &fixe.=1 and &prec2.=0 and &ajout_fixe2. ne " " %then %do;

        data temp_mod.select&no_modele._%eval(&nb_var.-1)a (drop = nomvar%eval(&nb_var.-3)0
nomvar%eval(&nb_var.-2)0 nomvar%eval(&nb_var.-1)0);
                set temp_mod.select&no_modele._%eval(&nb_var.-1)a (rename = (nomvar%eval(&nb_var.-
3) = nomvar%eval(&nb_var.-3)0

                        nomvar%eval(&nb_var.-2) = nomvar%eval(&nb_var.-2)0

                        nomvar%eval(&nb_var.-1) = nomvar%eval(&nb_var.-1)0));
                if nomvar%eval(&nb_var.-2)0 = &ajout_fixe2. then do;
                        nomvar%eval(&nb_var.-2) = nomvar%eval(&nb_var.-2)0;
                        if nomvar%eval(&nb_var.-3)0 = &ajout_fixe1. then do;
                                nomvar%eval(&nb_var.-3) = nomvar%eval(&nb_var.-3)0;
                                nomvar%eval(&nb_var.-1) = nomvar%eval(&nb_var.-1)0;
                        end;
                        else do;
                                nomvar%eval(&nb_var.-3) = nomvar%eval(&nb_var.-1)0;
                                nomvar%eval(&nb_var.-1) = nomvar%eval(&nb_var.-3)0;
                        end;
                end;
                else if nomvar%eval(&nb_var.-3)0 = &ajout_fixe2. then do;
                        nomvar%eval(&nb_var.-2) = nomvar%eval(&nb_var.-3)0;
                        if nomvar%eval(&nb_var.-2)0 = &ajout_fixe1. then do;
                                nomvar%eval(&nb_var.-3) = nomvar%eval(&nb_var.-2)0;
                                nomvar%eval(&nb_var.-1) = nomvar%eval(&nb_var.-1)0;
                        end;
                        else do;
                                nomvar%eval(&nb_var.-3) = nomvar%eval(&nb_var.-1)0;
                                nomvar%eval(&nb_var.-1) = nomvar%eval(&nb_var.-2)0;
                        end;
                end;
                else if nomvar%eval(&nb_var.-1)0 = &ajout_fixe2. then do;
                        nomvar%eval(&nb_var.-2) = nomvar%eval(&nb_var.-1)0;
                        if nomvar%eval(&nb_var.-3)0 = &ajout_fixe1. then do;
                                nomvar%eval(&nb_var.-3) = nomvar%eval(&nb_var.-3)0;
                                nomvar%eval(&nb_var.-1) = nomvar%eval(&nb_var.-2)0;
                        end;
                        else do;
                                nomvar%eval(&nb_var.-3) = nomvar%eval(&nb_var.-2)0;
                                nomvar%eval(&nb_var.-1) = nomvar%eval(&nb_var.-3)0;
                        end;
                end;
                else do;
                        nomvar%eval(&nb_var.-3) = nomvar%eval(&nb_var.-3)0;
                        nomvar%eval(&nb_var.-2) = nomvar%eval(&nb_var.-2)0;
                        nomvar%eval(&nb_var.-1) = nomvar%eval(&nb_var.-1)0;
                end;
        run;
%end;


proc sql;
        create table temp_mod.select&no_modele._&nb_var.a (where = (%do i=1 %to %eval(&nb_var.-2);
nomvar&nb_var. ne nomvar&i. and %end;

                                                                                nomvar&nb_var.
ne nomvar%eval(&nb_var.-1))
                                                                        drop = toutesvar0) as
                select *, (compbl(toutesvar0) || " " || compress(nomvar&nb_var.)) as toutesvar
```

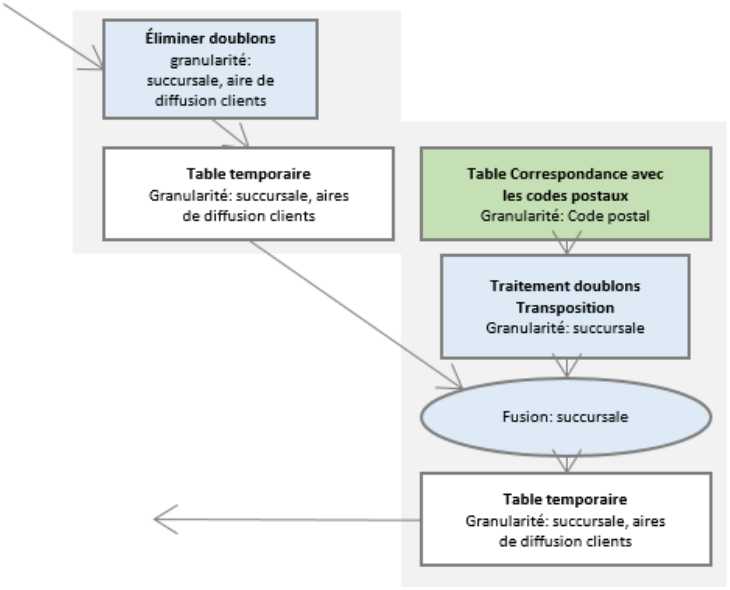
```
                from temp_mod.select&no_modele._%eval(&nb_var.-1)a (rename = (toutesvar =
toutesvar0)

        %if &fixe.=1 and &prec2.=1 %then %do;

        where = (nomvar%eval(&nb_var.-2) = &ajout_fixe1.)

        %end;

        %else %if &fixe.=1 and &prec2.=0 and &ajout_fixe2. ne " " %then %do;

        where = (nomvar%eval(&nb_var.-3) = &ajout_fixe1. and

                        nomvar%eval(&nb_var.-2) = &ajout_fixe2.)

        %end;

        ),
                        var_preselect_&no_modele. (rename = (nomvar = nomvar&nb_var.))
        ;
quit;


%if ((&fixe.=1 and &prec2.=1) or (&fixe.=1 and &prec2.=0 and &ajout_fixe2. ne " ")) %then %do;
data temp_mod.select&no_modele._&nb_var.a;
        set temp_mod.select&no_modele._&nb_var.a;
        if nomvar%eval(&nb_var.-1) < nomvar&nb_var. then toutesvar1 = compress(nomvar&nb_var.) || "
" || compress(nomvar%eval(&nb_var.-1));
        else if nomvar&nb_var. <= nomvar%eval(&nb_var.-1) then toutesvar1 =
compress(nomvar%eval(&nb_var.-1)) || " " || compress(nomvar&nb_var.);
run;

proc sort data = temp_mod.select&no_modele._&nb_var.a out = temp_mod.select&no_modele._&nb_var.a
nodupkey; by toutesvar1; run;

%end;

%if &fixe.=0 %then %do;

data temp_mod.select&no_modele._&nb_var.a;
        set temp_mod.select&no_modele._&nb_var.a;
        if nomvar%eval(&nb_var.-2) < nomvar%eval(&nb_var.-1) then do;
                if nomvar%eval(&nb_var.-2) < nomvar&nb_var. then do;
                        if nomvar%eval(&nb_var.-1) < nomvar&nb_var. then toutesvar1 =
compress(nomvar%eval(&nb_var.-2)) || " " || compress(nomvar%eval(&nb_var.-1)) || " " ||
compress(nomvar&nb_var.);
                        else toutesvar1 = compress(nomvar%eval(&nb_var.-2)) || " " ||
compress(nomvar&nb_var.) || " " || compress(nomvar%eval(&nb_var.-1));
                end;
                else toutesvar1 = compress(nomvar&nb_var.) || " " || compress(nomvar%eval(&nb_var.-
2)) || " " || compress(nomvar%eval(&nb_var.-1));
        end;
        else if nomvar%eval(&nb_var.-1) < nomvar%eval(&nb_var.-2) then do;
                if nomvar%eval(&nb_var.-1) < nomvar&nb_var. then do;
                        if nomvar%eval(&nb_var.-2) < nomvar&nb_var. then toutesvar1 =
compress(nomvar%eval(&nb_var.-1)) || " " || compress(nomvar%eval(&nb_var.-2)) || " " ||
compress(nomvar&nb_var.);
                        else toutesvar1 = compress(nomvar%eval(&nb_var.-1)) || " " ||
compress(nomvar&nb_var.) || " " || compress(nomvar%eval(&nb_var.-2));
                end;
                else toutesvar1 = compress(nomvar&nb_var.) || " " || compress(nomvar%eval(&nb_var.-
1)) || " " || compress(nomvar%eval(&nb_var.-2));
        end;
run;

proc sort data = temp_mod.select&no_modele._&nb_var.a out = temp_mod.select&no_modele._&nb_var.a
nodupkey; by toutesvar1; run;

%end;

%mend allposs2;

%macro inittotinfo;
data temp_mod.totinfocrit;
        length toutesvar $2000. Neg2LogLike parms aic aicc hqic bic caic 8.;
        toutesvar = ' ';
        Neg2LogLike = .;
```

```
        parms = .;
        aic = .;
        aicc = .;
        hqic = .;
        bic = .;
        caic = .;
run;
%mend inittotinfo;


*******************************************************************************************;
*** Modèles à 3 effets fixe - modèle 2;
*******************************************************************************************;

%let nb_var=3;

%allposs2(fixe=1, prec2=1, nb_var=&nb_var., ajout_fixe1="temps", ajout_fixe2=);
%inittotinfo;

ods select none;
data _null_;
        set temp_mod.select2_&nb_var.a;
        call execute ('%possibles(' || toutesvar || ')');
run;
ods select all;

%meilleur(aic, temp_mod.meilleurAIC2_&nb_var.effetsfixes);

*** Répliquer les lignes du modèle à 3 effets fixes pour tous les nombres de variables
suivant en modifiant les paramètres.;
```

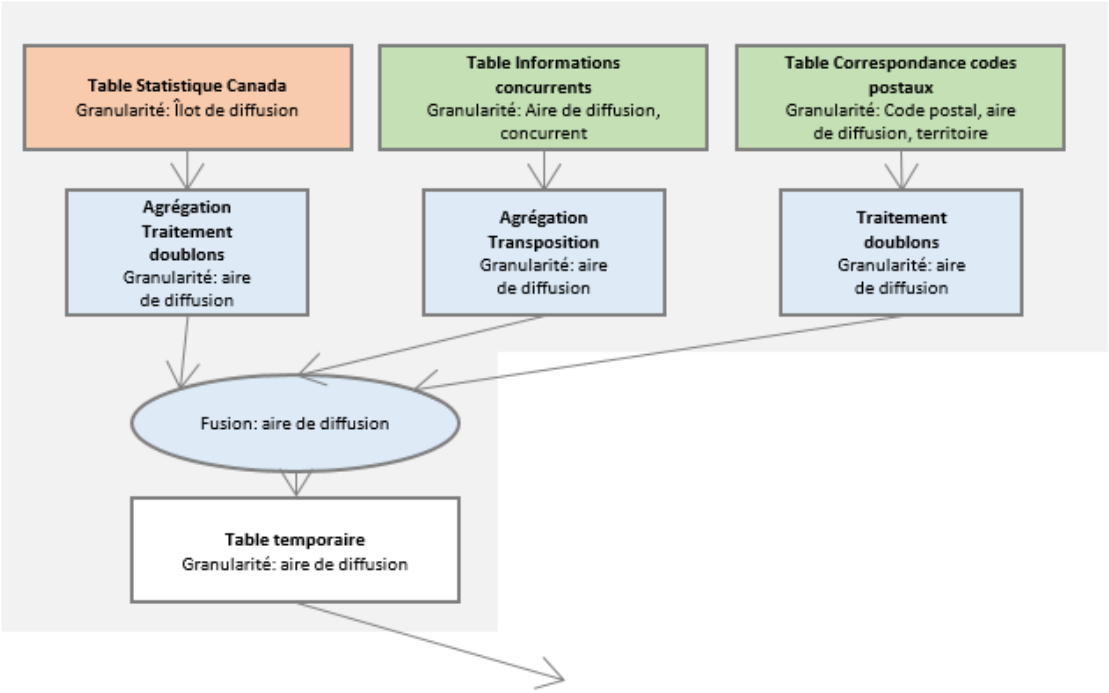
Annexe 2 : Vue scindée de la figure 2 (section 4.1)



Partie A



Partie B



Partie C

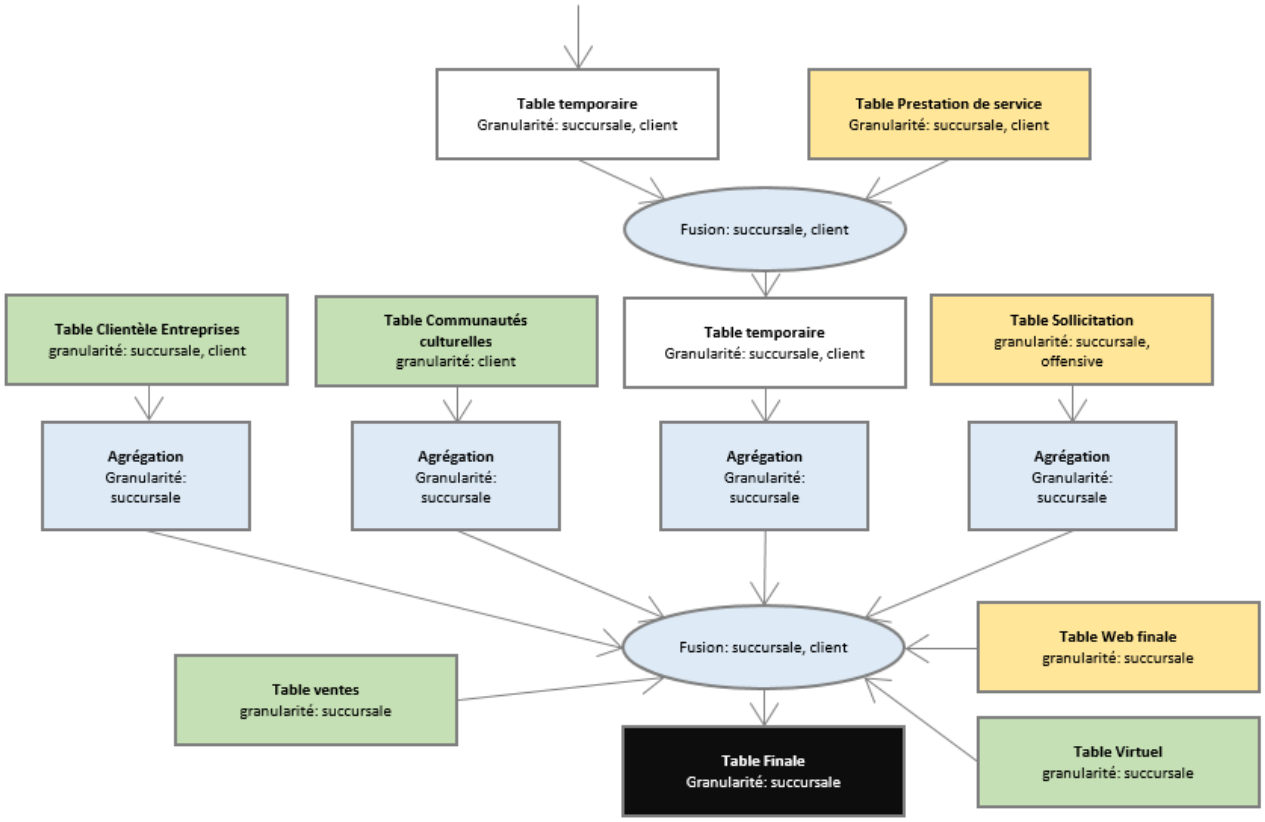


Table temporaire
Granularité: succursale, client
Table Prestation de service
Granularité: succursale, client
Fusion: succursale, client
Table Clientèle Entreprises
granularité: succursale, client
Table Communautés culturelles
granularité: client
Table temporaire
Granularité: succursale, client
Table Sollicitation
granularité: succursale, offensive
Agrégation
Granularité: succursale
Agrégation
Granularité: succursale
Agrégation
Granularité: succursale
Agrégation
Granularité: succursale
Fusion: succursale, client
Table Web finale
granularité: succursale
Table ventes
granularité: succursale
Table Finale
Granularité: succursale
Table Virtuel
granularité: succursale

Annexe 3 : Étapes permettant l'utilisation de la fonction VBA GoogleGeoCode

1. si ce n'est déjà fait, activer l'onglet développeur dans *Excel*. Dans le cas présent, cette étape avait déjà été effectuée;

2. dans l'onglet « développeur », cliquer sur « Visual Basic »;

3. cliquer sur « Insertion » et ensuite « Module »;

4. copier et coller le code VBA rendu disponible sur la page web mentionnée précédemment, en prenant soin de faire les modifications suivantes :

   Remplacer :

```
'define XML and HTTP components
Dim googleResult As New MSXML2.DOMDocument
Dim googleService As New MSXML2.XMLHTTP
Dim oNodes As MSXML2.IXMLDOMNodeList
Dim oNode As MSXML2.IXMLDOMNode
```

   Par :

```
'define XML and HTTP components
Dim googleResult As Object
Set googleResult = CreateObject("Msxml2.DOMDocument.6.0")
Dim googleService As Object
Set googleService = CreateObject("Msxml2.XMLHTTP.6.0")
Dim oNodes As MSXML2.IXMLDOMNodeList
Dim oNode As MSXML2.IXMLDOMNode
```

Le code VBA complet est présenté à la suite de l'étape 8;

5. dans le menu « Outils » de *Visual Basic*, cliquer sur « Références… ». Cocher la case associée à « Microsoft XML, v6.0 » et cliquer sur le bouton « OK »;

6. enregistrer le code inséré dans l'éditeur *Visual Basic* sous : GoogleGeoCode.xlam (en format « Macro complémentaire *Excel* »);

7. fermer tous les fichiers *Excel* ouverts et en ouvrir un nouveau. Dans le menu fichier, cliquer sur « Options » et ensuite sur « Compléments ». Cliquer sur le bouton « Atteindre », situé à côté de « Gérer : Compléments *Excel* »;

8. cocher la case de GoogleGeoCode.xlam et cliquer sur « OK ».

code VBA complet

(source : http://policeanalyst.com/using-the-google-geocoding-api-in-excel/ avec légères modifications)

```
Function GoogleGeocode(address As String) As String
Dim strAddress As String
Dim strQuery As String
Dim strLatitude As String
Dim strLongitude As String

strAddress = URLEncode(address)

'Assemble the query string
strQuery = "http://maps.googleapis.com/maps/api/geocode/xml?"
strQuery = strQuery & "address=" & strAddress
```

```
strQuery = strQuery & "&sensor=false"

'define XML and HTTP components
Dim googleResult As Object
Set googleResult = CreateObject("Msxml2.DOMDocument.6.0")
Dim googleService As Object
Set googleService = CreateObject("Msxml2.XMLHTTP.6.0")
Dim oNodes As MSXML2.IXMLDOMNodeList
Dim oNode As MSXML2.IXMLDOMNode

'create HTTP request to query URL - make sure to have
'that last "False" there for synchronous operation

googleService.Open "GET", strQuery, False
googleService.send
googleResult.LoadXML (googleService.responseText)

Set oNodes = googleResult.getElementsByTagName("geometry")

If oNodes.Length = 1 Then
        For Each oNode In oNodes
                strLatitude = oNode.ChildNodes(0).ChildNodes(0).Text
                strLongitude = oNode.ChildNodes(0).ChildNodes(1).Text
                GoogleGeocode = strLatitude & "," & strLongitude
        Next oNode
Else
        GoogleGeocode = "Not Found (try again, you may have done too many too fast)"
End If
End Function



Public Function URLEncode(StringVal As String, Optional SpaceAsPlus As Boolean = False)
As String
Dim StringLen As Long: StringLen = Len(StringVal)

If StringLen > 0 Then
        ReDim result(StringLen) As String
        Dim i As Long, CharCode As Integer
        Dim Char As String, Space As String

        If SpaceAsPlus Then Space = "+" Else Space = "%20"
```

```
        For i = 1 To StringLen
                Char = Mid$(StringVal, i, 1)
                CharCode = Asc(Char)

                Select Case CharCode
                Case 97 To 122, 65 To 90, 48 To 57, 45, 46, 95, 126
                        result(i) = Char
                Case 32
                        result(i) = Space
                Case 0 To 15
                        result(i) = "%0" & Hex(CharCode)
                Case Else
                        result(i) = "%" & Hex(CharCode)
                End Select
        Next i
        URLEncode = Join(result, "")
End If
End Function
```

Annexe 4 : Méthode développée pour l'obtention des latitudes et longitudes

1. Séparer le fichier de 4 504 villes distinctes (observations) en tables de données de 500 observations. Ces tables de données sont réunies à l'intérieur d'un seul fichier *Excel*, où chaque table est dans un onglet distinct.

2. Désactiver le calcul automatique du classeur *Excel*. Pour ce faire, par le menu « Formule », cliquer sur « Options de calcul » et sélectionner « Manuel ». Cette action est primordiale afin que l'application ne soumette pas constamment des requêtes inutiles à *Google*, qui conduiraient à un dépassement très rapide du quota journalier imposé par *Google*.

3. Pour une feuille de données (500 noms de villes), appliquer la fonction « GoogleGeocode » à chacun des noms de villes. Dans le menu « Formule », cliquer sur « Calculer la feuille ».

4. Pour chaque nom de ville où la fonction n'a pas retourné de valeur, soumettre de nouveau le calcul de la cellule. La touche « F2 » sur le clavier est très utile dans ce cas (remplace le « clic » dans la cellule).

5. Pour les noms de villes qui ne sont toujours pas associés à des coordonnées géographiques, tenter d'appliquer la fonction sur le nom de ville accompagné, dans l'ordre, des précisions géographiques Québec, Ontario, Canada et finalement US (États-Unis). Les noms de villes qui n'ont toujours pas été associés à des coordonnées géographiques sont ignorés.

6. Une fois le traitement sur la feuille terminé, copier et coller en valeurs la colonne où les informations de latitude et longitude ont été obtenues. Cette action permet d'éviter un nouveau calcul (nouvelle requête à *Google*) inutile.

7. Recommencer les quatre dernières étapes pour les autres feuilles de données. Les traitements ont dû être effectués en plusieurs étapes, sur plusieurs jours, étant donné la limite imposée par *Google*.

Annexe 5 : Tableau des variables

1. Variable dépendante

Nom du champ dans la source initiale	Description du champ de la source initiale	Source	Tables de données	Variables finales
Donnée confidentielle	Valeur suivi vente nette cumulative / Montant fin de période volume d'affaires	Entrepôts clientèle et comptes clients	Compte client	-Ratio ventes nettes / volume d'affaires de la succursale

2. Variables géographiques

Nom du champ dans la source initiale	Description du champ de la source initiale	Source	Tables de données	Variables finales
Donnée confidentielle	Numéro de la succursale	Entrepôts clientèle et comptes clients	Compte client	-Numéro de la succursale
Donnée confidentielle	Sous-secteur	Entrepôts clientèle et comptes clients	Compte client	-Sous-secteur
Donnée confidentielle	Secteur	Entrepôts clientèle et comptes clients	Compte client	-Secteur
Cod_ad_rec11 (+ Cod_ad_rec06)	Numéro de l'Aire de diffusion du recensement 2011 (du client ou de la succursale ou de l'institution financière concurrente)	Entrepôts clientèle et comptes clients	Correspondance avec les codes postaux Profil de la clientèle Informations concurrents	-% clients de la succursale dont l'aire de diffusion est celle de la succursale -Distance (moyenne, écart-type) entre la résidence principale des clients et la succursale
Cod_lati	Code de latitude (de la succursale)	Entrepôts clientèle et comptes clients	Sites physiques	
Cod_long	Code de longitude (de la succursale)	Entrepôts clientèle et comptes clients	Sites physiques	
ADidu (var7)	Aire de diffusion	Statistique Canada	Fichier des attributs géographiques	
ADlamx (var8)	Coordonnée x du point représentatif de l'AD en format Lambert	Statistique Canada	Fichier des attributs géographiques	
ADlamy (var9)	Coordonnée y du point représentatif de l'AD en format Lambert	Statistique Canada	Fichier des attributs géographiques	
ADlat (var10)	Latitude du point représentatif de l'AD, en degrés et décimales	Statistique Canada	Fichier des attributs géographiques	

ADlong (var11)	Longitude du point représentatif de l'AD, en degrés et décimales	Statistique Canada	Fichier des attributs géographiques	
CSSgenre (var28)	Regroupement des secteurs de recensement	Statistique Canada	Fichier des attributs géographiques	-% clients régions métropolitaines de recensement (1) -% clients agglomérations de recensement, au moins un secteur de recensement (2) -% clients agglomérations de recensement, sans secteur de recensement (3) -% clients extérieur de la région métropolitaine, une influence métropolitaine forte (4) -% clients extérieur de la région, influence métropolitaine faible (5) -% clients extérieur de la région métropolitaine, influence métropolitaine faible (6) -% clients extérieur de la région métropolitaine, aucune influence métropolitaine (7) -% clients territoire, extérieur des agglomérations de recensement (8)

CTROIORRgenre (var45)	Genre du centre de population et région rurale	Statistique Canada	Fichier des attributs géographiques	-% clients noyau à l'intérieur d'une région métropolitaine (1) -% clients Banlieue à l'intérieur d'une région métropolitaine (2) -% clients Région rurale à l'intérieur d'une région métropolitaine (3) -% clients Centre de population à l'extérieur d'une région métropolitaine (4) -% clients Région rurale à l'extérieur d'une région métropolitaine (5) -% clients Noyau secondaire à l'intérieur d'une région métropolitaine (6)
CTRPOPRRclasse (var46)	Permets la distinction entre les régions rurales, les petits centres de population, les moyens centres de population et les grands centres de population urbains.	Statistique Canada	Fichier des attributs géographiques	-% clients région rurale (1) -% clients Petit centre de population (2) -% clients Moyen centre de population (3) -% clients Grand centre de population urbain (4)
IDpop2011	Population de l'îlot de diffusion selon le Recensement de 2011	Statistique Canada	Fichier des attributs géographiques	-Population des AD de la succursale
IDtlog2011	Total des logements privés de l'îlot de diffusion selon le Recensement de 2011	Statistique Canada	Fichier des attributs géographiques	-Logements privés des AD de la succursale -Ratio nb de logements privés / population des AD

IDrh2011	Logements privés occupés par les résidents habituels de l'îlot de diffusion selon le Recensement de 2011	Statistique Canada	Fichier des attributs géographiques	-Logements privés occupés des AD de la succursale -Ratio nb de logements privés occupés / population des AD
Nom_infi_cncr	Nom de l'institution financière concurrente	Entrepôts clientèle et comptes clients	Informations concurrents	Nombre moyen et écart-type de chaque institution dans l'AD du client : - Nb inst1 - NB inst2 - NB inst3 - … - Nb total concurrent Nombre d'institutions financières concurrentes dans les AD des succursales : - Nb inst1 - Nb inst2 - NB inst3 - … - Nb total concurrent

3. Variables financières

Nom du champ dans la source initiale	Description du champ de la source initiale	Source	Tables de données	Variables finales
Donnée confidentielle	Indicateur de prêt étudiant en cours	Entrepôts clientèle et comptes clients	Crédit	-% des clients de la succursale ayant un prêt étudiant en cours
Donnée confidentielle	Montant autorisé de l'ensemble des prêts	Entrepôts clientèle et comptes clients	Profil de la clientèle	-$ prêts autorisés (moyenne, médiane, écart-type) -% clients prêts autorisés
Donnée confidentielle	Solde total des prêts	Entrepôts clientèle et comptes clients	Profil de la clientèle	-$ prêts (moyenne, médiane, écart-type) -% clients prêts
Donnée confidentielle	Valeur de la propriété financée	Entrepôts clientèle et comptes clients	Crédit	-Valeur propriétés financées (moyenne, médiane, écart-type)
Donnée confidentielle	Nombre d'hypothèques	Entrepôts clientèle et comptes clients	Profil de la clientèle	-Nombre moyen d'hypothèques détenues
Donnée confidentielle	Indicateur détention produit partenaire d'affaires 1	Entrepôts clientèle et comptes clients	Profil de la clientèle	-% détention produit partenaire d'affaires 1
Donnée confidentielle	Indicateur détention produit partenaire d'affaires 2	Entrepôts clientèle et comptes clients	Profil de la clientèle	-% détention produit partenaire d'affaires 2
Donnée confidentielle	Indicateur détention produit partenaire d'affaires 3	Entrepôts clientèle et comptes clients	Profil de la clientèle	-% détention produit partenaire d'affaires 3
Donnée confidentielle	Revenu brut du client	Entrepôts clientèle et comptes clients	Personnes	-Revenu brut (moyenne, médiane, écart-type)
Donnée confidentielle / Donnée confidentielle / Donnée confidentielle	Montant Écart Marché entre épargne estimée et épargne détenue dans l'entreprise/ Montant fin de période épargne détenue dans l'entreprise/	Entrepôts clientèle et comptes clients	Profil de la clientèle	-Écart marché (moyenne, médiane, écart-type) -Épargne Entreprise (moyenne, médiane, écart-type)

	Estimation de l'épargne marché			-Ratio écart marché moyen / épargne estimée marché moyen -Ratio épargne Entreprise / Épargne estimée marché -Épargne estimée (moyenne, médiane, écart-type)
Donnée confidentielle	Indicateur détention produit partenaire d'affaires 4	Entrepôts clientèle et comptes clients	Profil de la clientèle	-% détention produit partenaire d'affaires 4 -% détention produit partenaire d'affaires 4 affaires -% détention produit partenaire d'affaires 4 particuliers -% détention produit partenaire d'affaires 4 2 types
Donnée confidentielle	Solde des comptes Valeurs mobilières	Entrepôts clientèle et comptes clients	Valeurs mobilières	-$ Valeurs mobilières (moyenne, médiane, écart-type) -Ratio $ Valeurs mobilières / $ Épargne succursale -% clients Valeurs mobilières dans la succursale
Donnée confidentielle	Solde du compte Gestion discrétionnaire	Entrepôts clientèle et comptes clients	Profil de la clientèle	-$ Gestion discrétionnaire (moyenne, médiane, écart-type) -Ratio $ Gestion discrétionnaire / $ Épargne succursale

				-% clients Gestion discrétionnaire dans la succursale
Donnée confidentielle	Montant produit d'épargne 1	Entrepôts clientèle et comptes clients	Profil de la clientèle	-$ produit d'épargne 1 (moyenne, médiane, écart-type) -Ratio $ produit d'épargne 1 / $ Épargne succursale -% clients produit d'épargne 1 dans la succursale
Donnée confidentielle	Montant produit d'épargne 2	Entrepôts clientèle et comptes clients	Profil de la clientèle	-$ produit d'épargne 2 (moyenne, médiane, écart-type) -Ratio $ produit d'épargne 2 / $ Épargne succursale -% clients produit d'épargne 2
Donnée confidentielle	Montant produit d'épargne 3	Entrepôts clientèle et comptes clients	Profil de la clientèle	-$ produit d'épargne 3 (moyenne, médiane, écart-type) -Ratio $ produit d'épargne 3 / $ Épargne succursale -% clients produit d'épargne 3
Donnée confidentielle	Montant produit d'épargne 4	Entrepôts clientèle et comptes clients	Profil de la clientèle	-$ produit d'épargne 4 (moyenne, médiane, écart-type) -Ratio $ produit d'épargne 4 / $ épargne succursale -% clients produit d'épargne 4
Donnée confidentielle	Montant produit d'épargne 5	Entrepôts clientèle et comptes clients	Profil de la clientèle	-$ produit d'épargne 5 (moyenne, médiane, écart-type)

				-Ratio $ produit d'épargne 5 / $ épargne succursale -% clients produit d'épargne 5
Donnée confidentielle	Montant produit d'épargne 6	Entrepôts clientèle et comptes clients	Profil de la clientèle	-$ produit d'épargne 6 (moyenne, médiane, écart-type) -Ratio $ produit d'épargne 6 / $ épargne succursale
Donnée confidentielle	Montant produit d'épargne 7	Entrepôts clientèle et comptes clients	Profil de la clientèle	-$ produit d'épargne 7 (moyenne, médiane, écart-type) -Ratio $ produit d'épargne 7 / $ épargne succursale
Donnée confidentielle	Montant produit d'épargne 8	Entrepôts clientèle et comptes clients	Profil de la clientèle	-$ produit d'épargne 8 (moyenne, médiane, écart-type) -Ratio $ produit d'épargne 8 / $ épargne succursale -% clients produit d'épargne 8
Donnée confidentielle	Montant produit d'épargne 9	Entrepôts clientèle et comptes clients	Profil de la clientèle	-$ produit d'épargne 9 (moyenne, médiane, écart-type) -Ratio $ produit d'épargne 9 / $ épargne succursale -% clients produit d'épargne 9
Donnée confidentielle	Montant produit d'épargne 10	Entrepôts clientèle et comptes clients	Profil de la clientèle	-$ produit d'épargne 10 (moyenne, médiane, écart-type) -Ratio $ produit d'épargne 10 / $ épargne succursale

				-% clients produit d'épargne 10
Donnée confidentielle	Nombre de lignes de produits utilisées	Entrepôts clientèle et comptes clients	Profil de la clientèle	-Nombre moyen de lignes de produits utilisées
Donnée confidentielle	Volume d'affaires	Entrepôts clientèle et comptes clients	Profil de la clientèle	-Montant de volume d'affaires (moyenne, médiane, écart-type)
Donnée confidentielle	Ventes nettes en produits virtuels	Entrepôts clientèle et comptes clients	Compte client	-Montant total produits virtuels -Ratio ventes nettes produits virtuels / ventes nettes totales

4. Variables sociodémographiques

Nom du champ dans la source initiale	Description du champ de la source initiale	Source	Tables de données	Variables finales
Donnée confidentielle	Code de la communauté culturelle	Entrepôts clientèle et comptes clients	Communautés culturelles	-Nombre de communautés culturelles
Donnée confidentielle	Code de langue du client	Entrepôts clientèle et comptes clients	Personnes	-% Anglais -% Français -% Non défini
Donnée confidentielle	Indicateur de communauté culturelle	Entrepôts clientèle et comptes clients	Profil de la clientèle	-% communautés culturelles dans la succursale
Donnée confidentielle	Âge client	Entrepôts clientèle et comptes clients	Profil de la clientèle	-Âge (moyenne, médiane, écart-type) -% < 18 ans (1) -% entre 18 et 30 ans (2) -% entre 30 et 40 ans (3) -% entre 40 et 50 ans (4) -% entre 50 et 60 ans (5) -% entre 60 et 70 ans (6) -% 70 ans et plus (7) -% nd (999)
Donnée confidentielle	État civil du client	Entrepôts clientèle et comptes clients	Personnes	-% initial (21) -% Célibataire (22) -% Marié (23) -% Veuf (24) -% Séparé (25) -% Conjoint de fait (26) -% Non défini (99)
Donnée confidentielle	Statut d'emploi client	Entrepôts clientèle et comptes clients	Personnes	-% emploi permanent
Donnée confidentielle	Statut immigrant du client	Entrepôts clientèle et comptes clients	Personnes	-% immigrant
Donnée confidentielle	Cycle de vie financier	Entrepôts clientèle et comptes clients	Profil de la clientèle	-% moins de 18 ans (0) -% Étudiants (1)

				-% Jeunes travailleurs (2) -% Accédant Propriété (3) -% Nouveaux propriétaires (4) -% Projets divers (5) -% préparation retraite (6) -% retraite (8) -% valeur manquante (999)
Donnée confidentielle	Indicateur chef d'entreprise	Entrepôts clientèle et comptes clients	Profil de la clientèle	-% chef d'entreprise
Donnée confidentielle	Indicateur travailleur autonome	Entrepôts clientèle et comptes clients	Profil de la clientèle	-% travailleurs autonomes
Donnée confidentielle	Niveau de richesse	Entrepôts clientèle et comptes clients	Profil de la clientèle	-% Moins de 18 ans (0) -% Utilisateurs (1) -% Bâtisseurs (2) -% Accumulateurs (3) -% Aisé 1 (41) -% Aisé 2 (42) -% Fortunés (5) -% valeurs manquantes (999)
Donnée confidentielle	Clients propriétaires	Entrepôts clientèle et comptes clients	Profil de la clientèle	-% clients propriétaires
Donnée confidentielle	Propriétaires d'immeubles locatifs	Entrepôts clientèle et comptes clients	Indicateurs de segmentation	-% non propriétaire immeuble locatif (0) -% Propriétaire d'immeubles multilocatifs (1) -% Investisseurs immobiliers (2)
Donnée confidentielle	Client Demuni	Entrepôts clientèle et comptes clients	Indicateurs de segmentation	-% clients démunis
Donnée confidentielle	Code identifiant des professionnels faisant partie de l'offre distinction	Entrepôts clientèle et comptes clients	Indicateurs de segmentation	-% Offre distinction

Donnée confidentielle	Code d'occupation du client	Entrepôts clientèle et comptes clients	Profil de la clientèle	-% Personnel-cadre (1) -% Professions libérales et techniciens (2) -% Employés de bureau (3) -% Vendeurs (4) -% Travailleurs de services et des activités récréatives (5) -% Travailleurs des transports et communications (6) -% Travailleurs agricoles forestiers mines (7) -% Ouvriers spécialisés (8) -% Manœuvre ou journaliers (9) -% Maîtresse de maison / femme au foyer (10) -% Enfant (11) -% Étudiant niveau postsecondaire (13) -% Sans occupation ou retraité (14) -% Non défini (99)
Donnée confidentielle	Type résidence	Entrepôts clientèle et comptes clients	Personnes	-% Aucun -% Propriétaire -% locataire -% chambreur -% parents -% autres -% non disponible
Donnée confidentielle	Décès du client	Entrepôts clientèle et comptes clients	Personnes	-% non décédés (0) -% décédés (1) -% Succession (37)

				-% non défini (99)
Donnée confidentielle	Code de regroupement professionnel	Entrepôts clientèle et comptes clients	Personnes	
Donnée confidentielle	Code de sexe	Entrepôts clientèle et comptes clients	Personnes	-% hommes -% femmes -% non disponible
Donnée confidentielle	Code type d'entreprises	Entrepôts clientèle et comptes clients	Clientèle Entreprises	-% pas d'info -% micro entreprises – Travailleur autonome -% Exploitation imm. Résidentiels -% Très petite entreprise -% Petite entreprise -% Moyenne entreprise -% Grande entreprise -Ratios nb du type d'entreprise / nb clients entreprises (plusieurs ratios)
Donnée confidentielle	Code type entreprise légal	Entrepôts clientèle et comptes clients	Clientèle Entreprises	-% enregistré -% Incorporé -% Limité -% Société en nom collectif -% aucune forme identifiée -Ratios nb du type légal / nb clients entreprises (plusieurs ratios)
Donnée confidentielle	Nombre de client	Entrepôts clientèle et comptes clients	Clientèle Entreprises	-% clients entreprises
Donnée confidentielle	Code type de financement	Entrepôts clientèle et comptes clients	Clientèle Entreprises	-% agricole -% Commercial / Industriel -% Institutionnel -% Multilocatif -% Aucun secteur identifié

				-Ratios nb du type de financement / nb clients entreprises (plusieurs ratios)
Donnée confidentielle	Indicateur entreprise « division entreprise »	Entrepôts clientèle et comptes clients	Clientèle Entreprises	-% division entreprises -Ratio nb division entreprises / nb clients entreprises
Donnée confidentielle	Montant du chiffre d'affaire estimé	Entrepôts clientèle et comptes clients	Clientèle Entreprises	-Montant moyen chiffre d'affaires pour les entreprises -Écart-type chiffre d'affaires pour entreprises
Donnée confidentielle	Montant revenu net d'affaires mensuel	Entrepôts clientèle et comptes clients	Clientèle Entreprises	-Montant moyen revenu net mens pour les entreprises -Écart-type montant revenu net mens pour les entreprises
Donnée confidentielle	Nombre d'employés de l'entreprise	Entrepôts clientèle et comptes clients	Clientèle Entreprises	-Nombre moyen d'employés pour les entreprises -Écart-type nombre d'employés pour les entreprises
NOTEMAT	Indice de défavorisation – note factorielle matérielle	INSPQ	TableEquivalenceComplete Quebec2006.xls	-Note composante matérielle (moyenne, médiane, écart-type)
NOTESOC	Indice de défavorisation – note factorielle sociale	INSPQ	TableEquivalenceComplete Quebec2006.xls	-Note composante sociale (moyenne, médiane, écart-type)
QuintMat	Variations nationales (provinciales) de l'indice de défavorisation – Quintile matériel (1 à 5)	INSPQ	TableEquivalenceComplete Quebec2006.xls	-Quintile matériel (moyenne, médiane, écart-type)
QuintSoc	Variations nationales (provinciales) de l'indice de	INSPQ	TableEquivalenceComplete Quebec2006.xls	-Quintile social (moyenne, médiane, écart-type)

	défavorisation – Quintile social (1 à 5)			
CentMat	Indice de défavorisation – centile matériel (1 à 100)	INSPQ	TableEquivalenceCompleteQuebec2006.xls	-Centile matériel (moyenne, médiane, écart-type)
CentSoc	Indice de défavorisation – centile social (1 à 100)	INSPQ	TableEquivalenceCompleteQuebec2006.xls	-Centile social (moyenne, médiane, écart-type)
Quintmatrc	Variations canadiennes locales (RC) de l'indice de défavorisation – Quintile matériel (1 à 5)	INSPQ	TableEquivalenceCompleteCanada2006.xls	-Quintile local rc matériel (moyenne, médiane, écart-type)
Quintsocrc	Variations canadiennes locales (RC) de l'indice de défavorisation – Quintile social (1 à 5)	INSPQ	TableEquivalenceCompleteCanada2006.xls	-Quintile local rc social (moyenne, médiane, écart-type)
Quintmatzone	Variations canadiennes locales (Zone) de l'indice de défavorisation – Quintile matériel (1 à 5)	INSPQ	TableEquivalenceCompleteCanada2006.xls	-Quintile local zone matériel (moyenne, médiane, écart-type)
Quintsoczone	Variations canadiennes locales (Zone) de l'indice de défavorisation – Quintile matériel (1 à 5)	INSPQ	TableEquivalenceCompleteCanada2006.xls	-Quintile local zone social (moyenne, médiane, écart-type)

5. Variables Comportement

Nom du champ dans la source initiale	Description du champ de la source initiale	Source	Tables de données	Variables finales
Donnée confidentielle	Nombre d'années depuis l'ouverture du premier compte	Entrepôts clientèle et comptes clients	Profil de la clientèle	-Ancienneté succursale (moyenne, médiane, écart-type)
Donnée confidentielle	Indicateur d'adhésion aux services en ligne	Entrepôts clientèle et comptes clients	Profil de la clientèle	-% adhésion services en ligne
Donnée confidentielle	Équipe du conseiller de la personne	Entrepôts clientèle et comptes clients	Profil de la clientèle	-% équipe 1 (1) -% équipe 2 (2) -% équipe 3 (3) -% non défini (9)
Donnée confidentielle	Code d'utilisation des services en ligne	Entrepôts clientèle et comptes clients	Profil de la clientèle	-% transactions en ligne -Moyenne indice utilisation (où 1 = aucune utilisation, 2 = téléphone info seul, 3 = téléphone transactions, 4 = Internet info seul, 5 = internet transactions) -% 1 à % 5 -% valeur manquante (999)
Donnée confidentielle	Institution transactionnelle principale	Entrepôts clientèle et comptes clients	Profil de la clientèle	-% institution transactionnelle principale
Donnée confidentielle	Relation d'affaires	Entrepôts clientèle et comptes clients	Profil de la clientèle	-% principaux concurrents -% mixtes épargne et financement -% Mixtes financement -% Mixtes épargne -% Princ. – Faible potentiel -% Princ. – Fort potentiel

6. Variables Prestation de service et sollicitation

Nom du champ dans la source initiale	Description du champ de la source initiale	Source	Tables de données	Variables finales
NbProjet	Nombre de projets dans l'outil de prestation de service	Prestation de service	Prestation de service	-% clients ayant au moins un projet complété -Nb moyen de projet
Plan financier	Indicateur Plan financier avec option complété/non complété	Prestation de service	Prestation de service	-% clients plan financier ouvert -% clients plan financier peu complété -% clients plan financier avancés -$ comptes projet (moyenne, médiane, écart-type)
Potentiel de rapatriement	Potentiel de rapatriement identifié dans l'outil de prestation de service	Prestation de service	Prestation de service	-Potentiel de rapatriement identifié chez les clients ayant un plan financier (moyenne, médiane, écart-type) -% rapatrié
Sollicitation	Indicateur adhésion plan de sollicitation	Sollicitation	Sollicitation	-Indicateurs adhésion E, O ou R -Indicateurs offre cible E, O ou R

7. Variables Web

Nom du champ dans la source initiale	Description du champ de la source initiale	Source	Tables de données	Variables finales
Visites	Nombre de visites sur les pages Épargne	GA	GA	-Ratio nb visites / nb clients, par groupes d'heures (1, 2, 3) -% visites, par groupes d'heures (1, 2, 3)
Nouveaux utilisateurs	Nombre de nouveaux utilisateurs sur les pages Épargne	GA	GA	-Nb nouveaux utilisateurs, par groupe d'heures (1, 2, 3) -Ratios Nb nouveaux utilisateurs / nb clients, par groupe d'heures (1, 2, 3)
Pages vues	Nombre de pages vues sur les pages Épargne	GA	GA	-Nb pages vues / session, par groupe d'heures (1, 2, 3)
Durée	Durées moyennes des visites	GA	GA	-Durée moyenne / session, par groupe d'heures (1, 2, 3)
Ville	Sur les pages Épargne	GA	GA	-Nombre de villes GA associées à la succursale

8. Autres variables

Nom du champ dans la source initiale	Description du champ de la source initiale	Source	Tables de données	Variables finales
Temps	Position des différents trimestres dans l'étude (1 à 7)			-Temps
Trimestre	Trimestre dans l'année (1 à 4)			-Trimestre

Annexe 6 : Détails des étapes de sélection de modèles et de variables

Données des entrepôts de données internes

Nombre de facteurs fixes	Nombre de possibilités	AIC	BIC	Variables fixées ajoutées	Variables sélectionnées
1	43	-13150,8598	-13034,27668		-temps
2	903	-13201,25163	-13080,78241		-temps - Montant total de ventes nettes en produits virtuels
3	861	-13210,12123	-13085,7659	-temps	-Montant médian autorisé pour l'ensemble des prêts - Montant total de ventes nettes en produits virtuels
4	820	-13214,9124	-13086,67097	-Montant total de ventes nettes en produits virtuels	-Proportion de clients dont l'état civil est «marié» -Montant médian autorisé pour l'ensemble des prêts
5	780	-13222,2597	-13090,13217	-Montant médian autorisé pour l'ensemble des prêts	-Proportion des clients appartenant à une communauté culturelle -Proportion de clients dont l'état civil est «marié»

6	741	-13233,13494	-13097,12129	-Proportion de clients dont l'état civil est «marié»	-Proportion de clients locataires -Proportion des clients appartenant à une communauté culturelle
7	703	-13239,12972	-13099,22997	-Proportion des clients appartenant à une communauté culturelle	-Écart-type du montant détenu dans le produit d'épargne 4 -Proportion de clients locataires
8	666	-13243,85658	-13100,07073	-Proportion de clients locataires	-Écart-type du montant détenu dans le produit d'épargne 4 -Proportion de clients détenant le produit d'épargne 3
9	630	-13247,51449	-13099,84254	-Écart-type du montant détenu dans le produit d'épargne 4	-Proportion des clients dont la relation d'affaires est «mixte financement» -Proportion de clients détenant le produit d'épargne 3
10	595	-13253,15237	-13101,59432	-Proportion de clients détenant le produit d'épargne 3	-Proportion de clients dont le code d'occupation est «profession libérale ou technicien» -Proportion des clients dont la relation d'affaires est «mixte financement»
11	561	-13257,95934	-13102,51518	-Proportion des clients dont la relation d'affaires est «mixte financement»	-Proportion de clients dont le code d'occupation est «profession libérale ou technicien» -Proportion de clients dont l'état civil est «célibataire»

12	528	-13259,387	-13100,05674	-Proportion de clients dont le code d'occupation est «profession libérale ou technicien»	-Proportion de clients dont l'état civil est «célibataire» -Proportion de clients dont l'état civil est «initial»
13	496	-13267,08654	-13103,87017	-Proportion de clients dont l'état civil est «célibataire»	-Proportion de clients habitant chez leurs parents -Proportion de clients dont l'état civil est «initial»
14	465	-13272,19168	-13105,0892	-Proportion de clients dont l'état civil est «initial»	-Ratio montant d'écart d'épargne estimée totale marché / montant total estimé marché -Proportion de clients habitant chez leurs parents
15	435	-13272,57345	-13101,58487	-Proportion de clients habitant chez leurs parents	-Ratio montant d'écart d'épargne estimée totale marché / montant total estimé marché -Proportion de clients détenant le produit d'épargne 1
16	406	-13273,23633	-13098,36164	-Ratio montant d'écart d'épargne estimée totale marché / montant total estimé marché	-Écart-type du montant détenu dans le produit d'épargne 8 -Proportion de clients détenant le produit d'épargne 1
17	378	-13273,6965	-13094,93571	-Proportion de clients détenant le produit d'épargne 1	-Écart-type du montant détenu dans le produit d'épargne 8 - Proportion de clients dont le niveau de richesse est «Aisé 1»

18	351	-13274,12068	-13091,47379	-Écart-type du montant détenu dans le produit d'épargne 8	-Proportion de clients chambreurs -Proportion de clients dont le niveau de richesse est «Aisé 1»
19	325	-13273,48063	-13086,94764	-Proportion de clients dont le niveau de richesse est «Aisé 1»	-Proportion de clients chambreurs -Proportion de clients démunis
20	300	-13273,43056	-13083,01146	-Proportion de clients chambreurs	-Proportion des clients utilisant les services par téléphone pour de l'information seulement -Proportion de clients démunis
21	276	-13272,70465	-13078,39945	-Proportion de clients démunis	-Proportion des clients utilisant les services par téléphone pour de l'information seulement -Montant moyen du chiffre d'affaires des entreprises
22	253	-13271,70398	-13073,51267	-Proportion des clients utilisant les services par téléphone pour de l'information seulement	-Proportion de clients - cycle de vie «retraite» -Proportion de clients détenteurs d'une carte de crédit pour «Particuliers» de l'entreprise

Données provenant de multiples sources

Nombre de facteurs fixes	Nombre de possibilités	AIC	BIC	Varialbles fixées	Variables sélectionnées
1		-13150,8598	-13034,27668		-temps
2	703	-13201,25163	-13080,78241		-temps -Montant total de ventes nettes en produits virtuels
3	666	-13221,57756	-13097,22223	-temps	-Durée moyenne des sessions dans le groupe d'heures 2 -Montant total de ventes nettes en produits virtuels
4	630	-13231,36558	-13103,12415	-Montant total de ventes nettes en produits virtuels	-Durée moyenne des sessions dans le groupe d'heures 2 -Montant médian autorisé pour l'ensemble des prêts
5	595	-13234,70104	-13102,57351	-Durée moyenne des sessions dans le groupe d'heures 2	-Ratio nombre de sessions dans le groupe d'heures 3 / nombre de sessions totales -Montant médian autorisé pour l'ensemble des prêts
6	561	-13237,0326	-13101,01896	-Montant médian autorisé pour l'ensemble des prêts	-Proportion de clients dont l'état civil est «célibataire» -Proportion de clients dont l'état civil est «initial»
7	5984	-13241,29017	-13101,39042		-Proportion de clients habitant à l'extérieur d'une région métropolitaine, aucune influence métropolitaine -Proportion de clients dont l'état civil est «célibataire» -Proportion de clients dont l'état civil est «initial»

8	496	-13247,91048	-13104,12463	-Proportion de clients dont l'état civil est «célibataire» -Proportion de clients dont l'état civil est «initial»	-Proportion des clients appartenant à une communauté culturelle -Proportion de clients détenant le produit d'épargne 3
9	4960	-13253,03515	-13105,3632		-Proportion de clients habitant à l'extérieur d'une région métropolitaine, aucune influence métropolitaine -Proportion de clients détenant le produit d'épargne 3 -Proportion des clients appartenant à une communauté culturelle
10	435	-13260,10908	-13108,55102	-Proportion de clients détenant le produit d'épargne 3 -Proportion des clients appartenant à une communauté culturelle	-Écart-type du montant détenu dans le produit d'épargne 4 -Proportion de clients habitant à l'extérieur d'une région métropolitaine, aucune influence métropolitaine
11	406	-13264,02152	-13108,57736	-Proportion de clients habitant à l'extérieur d'une région métropolitaine, aucune influence métropolitaine	-Ratio nombre de sessions dans le groupe d'heures 3 / nombre de sessions totales -Écart-type du montant détenu dans le produit d'épargne 4
12	378	-13267,65309	-13108,32282	-Écart-type du montant détenu dans le produit d'épargne 4	-Écart-type du nombre de succursales du concurrent 5 dans l'aire de diffusion du client -Nombre moyen de succursales du concurrent 5 dans l'aire de diffusion du client
13	3276	-13271,57483	-13108,35846		-Ratio nombre de sessions dans le groupe d'heures 3 / nombre de sessions totales -Nombre moyen de succursales du concurrent 5 dans l'aire de diffusion du client -Écart-type du nombre de succursales du concurrent 5 dans l'aire de diffusion du client

14	325	-13274,00071	-13106,89824	-Nombre moyen de succursales du concurrent 5 dans l'aire de diffusion du client -Écart-type du nombre de succursales du concurrent 5 dans l'aire de diffusion du client	-Ratio nombre de sessions dans le groupe d'heures 3 / nombre de sessions totales -Proportion de clients dont le code d'occupation est «profession libérale ou technicien»
15	300	-13275,89839	-13104,90981	-Ratio nombre de sessions dans le groupe d'heures 3 / nombre de sessions totales	-Proportion de clients habitant chez leurs parents -Proportion de clients locataires
16	2300	-13282,94984	-13108,07516		-Proportion de clients locataires -Proportion de clients dont le code d'occupation est «profession libérale ou technicien» -Proportion de clients habitant chez leurs parents
17	253	-13285,86017	-13107,09939	-Proportion de clients habitant chez leurs parents -Proportion de clients locataires	-Ratio nombre de logements privés occupés / population des aires de diffusion -Proportion de clients dont le code d'occupation est «profession libérale ou technicien»
18	231	-13287,51845	-13104,87156	-Proportion de clients dont le code d'occupation est «profession libérale ou technicien»	-Ratio nombre de logements privés occupés / population des aires de diffusion -Proportion de clients dont le code d'occupation est «vendeur»
19	210	-13289,303	-13102,77	-Ratio nombre de logements privés occupés / population des aires de diffusion	-Proportion de clients dont le code d'occupation est «vendeur» -Proportion de clients dont l'état civil est «séparé»
20	190	-13290,76282	-13100,34372	-Proportion de clients dont le code d'occupation est «vendeur»	-Distance moyenne entre la résidence principale des clients et la succursale -Proportion de clients dont l'état civil est «séparé»
21	171	-13292,31217	-13098,00697	-Proportion de clients dont l'état civil est «séparé»	-Distance moyenne entre la résidence principale des clients et la succursale -Proportion de clients dont le niveau de richesse est «bâtisseur»

22	153	-13293,70991	-13095,5186	-Distance moyenne entre la résidence principale des clients et la succursale	-Proportion de clients chambreurs -Proportion de clients détenant un produit d'épargne 4
23	816	-13294,73039	-13092,65298		-Proportion de clients dont le niveau de richesse est «bâtisseur» -Proportion de clients détenant un produit d'épargne 4 -Proportion de clients chambreurs
24	120	-13295,37519	-13089,41168	-Proportion de clients détenant un produit d'épargne 4 -Proportion de clients chambreurs	-Ratio nombre de sessions dans le groupe d'heures 2 / nombre de sessions totales -Offre d'adhésion au plan de sollicitation de la cible E
25	560	-13296,21597	-13086,36635		-Offre d'adhésion au plan de sollicitation de la cible E -Proportion de clients dont le niveau de richesse est «bâtisseur» -Ratio nombre de sessions dans le groupe d'heures 2 / nombre de sessions totales
26	91	-13296,75918	-13083,02346	-Offre d'adhésion au plan de sollicitation de la cible E -Ratio nombre de sessions dans le groupe d'heures 2 / nombre de sessions totales	-Proportion de clients dont le niveau de richesse est «bâtisseur» -Proportion de clients dont l'état civil est «conjoint de fait»
27	78	-13297,53522	-13079,9134	-Proportion de clients dont le niveau de richesse est «bâtisseur»	-Proportion d'investisseurs immobiliers -Proportion de clients dont l'état civil est «conjoint de fait»
28	66	-13298,06981	-13076,56188	-Proportion de clients dont l'état civil est «conjoint de fait»	-Proportion d'investisseurs immobiliers -Nombre moyen de succursales du concurrent 6 dans l'aire de diffusion du client
29	55	-13298,27106	-13072,87703	-Proportion d'investisseurs immobiliers	-Proportion de clients dont le niveau de richesse est «Aisé 1» -Nombre moyen de succursales du concurrent 6 dans l'aire de diffusion du client

30	45	-13298,33176	-13069,05162	-Nombre moyen de succursales du concurrent 6 dans l'aire de diffusion du client	-Proportion de clients dont le niveau de richesse est «Aisé 1» -Proportion de clients attitrés à l'équipe 3
31	36	-13297,36122	-13064,19497	-Proportion de clients dont le niveau de richesse est «Aisé 1»	-Proportion de clients dont le code d'occupation est «ouvrier spécialisé» -Proportion de clients attitrés à l'équipe 3
32	28	-13296,16849	-13059,11614	-Proportion de clients attitrés à l'équipe 3	-Proportion de clients dont le code d'occupation est «ouvrier spécialisé» -Proportion de clients faisant affaire avec la filiale de gestion discrétionnaire
33	21	-13294,70583	-13053,76738	-Proportion de clients dont le code d'occupation est «ouvrier spécialisé»	-Montant moyen détenu dans le produit d'épargne 4 -Proportion de clients faisant affaire avec la filiale de gestion discrétionnaire

Annexe 7 : Détails des étapes pour l'ajout d'interactions et de termes d'ordre supérieurs

Modèle 1 : Données des entrepôts internes

Nombre d'effets supplémentaires	BIC	Effet
1	-13129.61175	Montant médian autorisé pour l'ensemble des prêts*Ratio montant d'écart d'épargne estimée totale marché / montant total estimé marché
2	-13143.92678	Proportion de clients locataires*Proportion de clients habitant chez leurs parents
3	-13152.44535	Ratio montant d'écart d'épargne estimée totale marché / montant total estimé marché*Proportion des clients dont la relation d'affaires est «mixte financement»
4	-13160.7732	Montant médian autorisé pour l'ensemble des prêts*Proportion de clients dont l'état civil est «initial»
5	-13167.84413	Ratio montant d'écart d'épargne estimée totale marché / montant total estimé marché*Proportion de clients dont l'état civil est «marié»
6	-13174.43085	Proportion des clients dont la relation d'affaires est «mixte financement»*Proportion des clients dont la relation d'affaires est «mixte financement»
7	-13180.17236	temps*sqrt(Écart-type du montant détenu dans le produit d'épargne 4)
8	-13185.75223	Sqrt(Écart-type du montant détenu dans le produit d'épargne 4)*sqrt(Écart-type du montant détenu dans le produit d'épargne 4)
9	-13189.27545	Proportion de clients locataires*Proportion de clients dont le code d'occupation est «profession libérale ou technicien»
10	-13192.5873	Proportion de clients dont l'état civil est «marié»*Proportion de clients détenant le produit d'épargne 3

11	-13192.91026	temps*temps
12	-13192.78165	temps*Proportion de clients habitant chez leurs parents
13	-13194.86903	temps*Proportion de clients dont le code d'occupation est «profession libérale ou technicien»
14	-13194.63909	Proportion de clients habitant chez leurs parents*Proportion de clients détenant le produit d'épargne 3
15	-13194.10176	Proportion de clients dont l'état civil est «initial»*Proportion de clients dont l'état civil est «initial»
16	-13196.96061	Proportion de clients détenant le produit d'épargne 3*Proportion de clients dont l'état civil est «célibataire»
17	-13198.81607	Sqrt(Écart-type du montant détenu dans le produit d'épargne 4)*Proportion de clients habitant chez leurs parents
18	-13198.61801	Proportion de clients locataires*Proportion de clients détenant le produit d'épargne 3
19	-13198.54218	Proportion de clients dont l'état civil est «célibataire»*Proportion de clients habitant chez leurs parents
20	-13197.65856	Sqrt(Écart-type du montant détenu dans le produit d'épargne 4)*Proportion de clients dont l'état civil est «marié»
21	-13196.44755	Proportion de clients habitant chez leurs parents* Proportion de clients habitant chez leurs parents

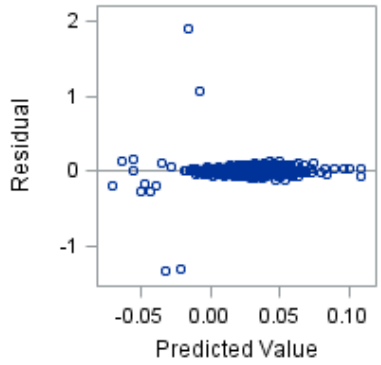
Modèle 2 : Données de sources multiples

Nombre d'effets supplémentaires	BIC	Effet
1	-13119.51604	temps*log(Écart-type du montant détenu dans le produit d'épargne 4)
2	-13124.07814	sqrt(Montant médian autorisé pour l'ensemble des prêts)*log(Écart-type du montant détenu dans le produit d'épargne 4)
3	-13127.50466	sqrt(Proportion de clients détenant le produit d'épargne 3)*sqrt(Proportion de clients habitant à l'extérieur d'une région métropolitaine, aucune influence métropolitaine)
4	-13126.85016	Proportion de clients dont l'état civil est «célibataire»*log(Écart-type du montant détenu dans le produit d'épargne 4)
5	-13127.23901	Proportion de clients dont l'état civil est «célibataire»*Proportion des clients appartenant à une communauté culturelle
6	-13127.64441	sqrt(Montant total de ventes nettes en produits virtuels)*sqrt(Proportion de clients habitant à l'extérieur d'une région métropolitaine, aucune influence métropolitaine)
7	-13126.55234	temps*sqrt(Montant médian autorisé pour l'ensemble des prêts)
8	-13124.47844	Proportion de clients dont l'état civil est «initial»*sqrt(Proportion de clients habitant à l'extérieur d'une région métropolitaine, aucune influence métropolitaine)
9	-13123.91341	Proportion de clients dont l'état civil est «initial»*sqrt(Proportion de clients détenant le produit d'épargne 3)

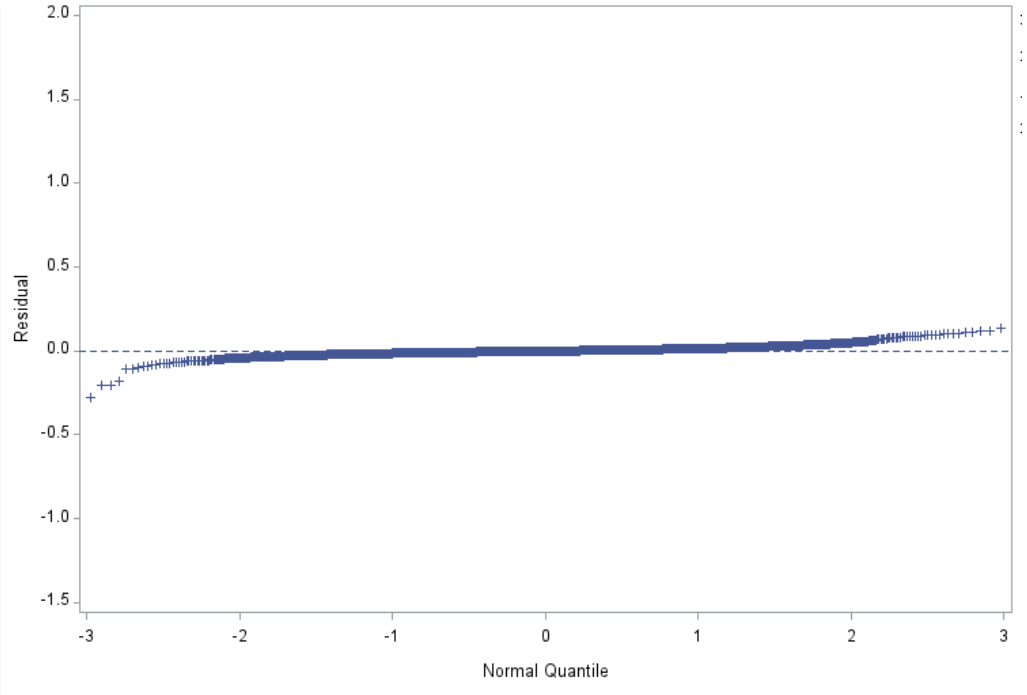
Annexe 8 : Graphiques pour la validation des modèles

Validation du modèle 1

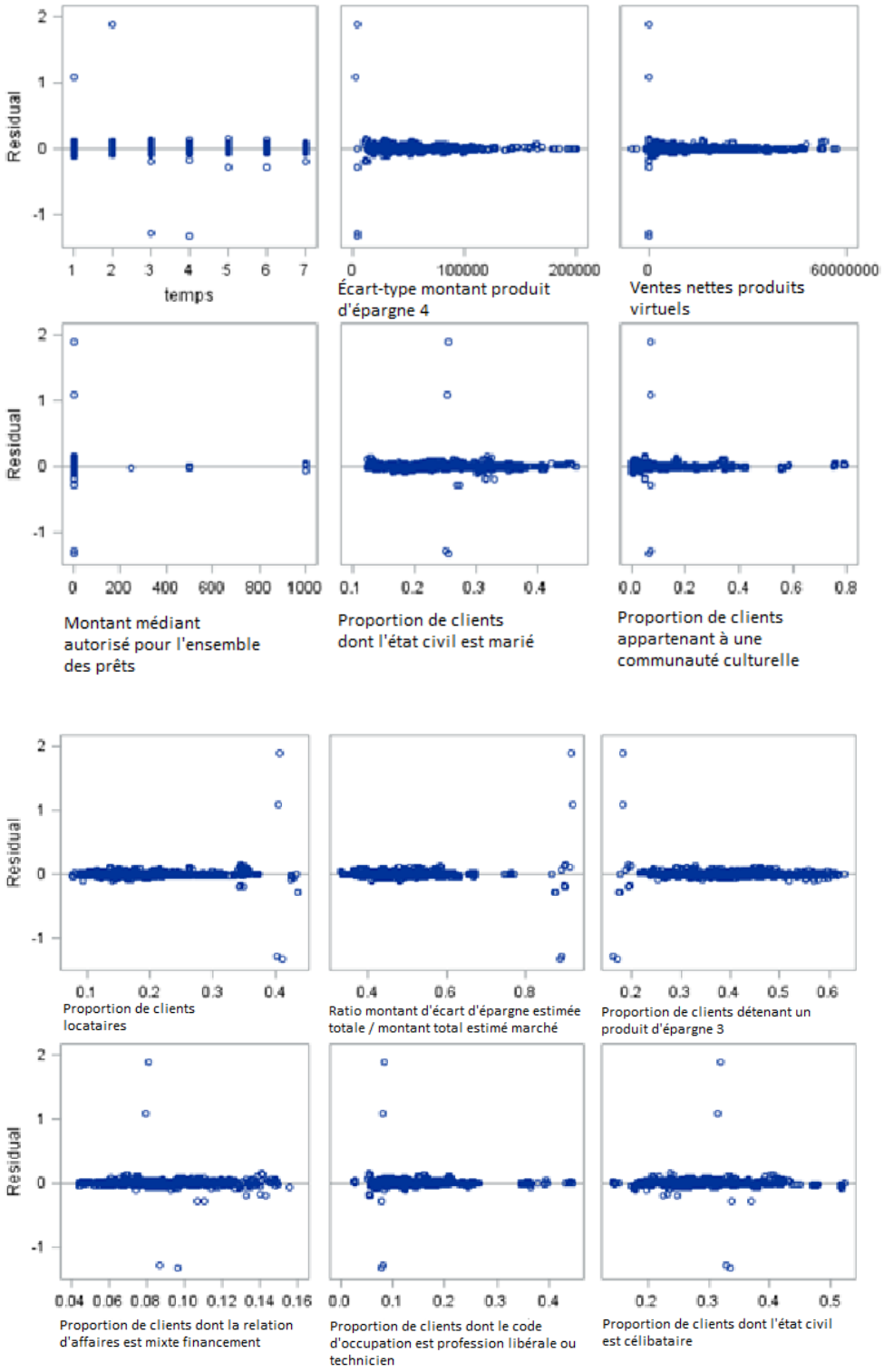
Résidus en fonction des valeurs prédites

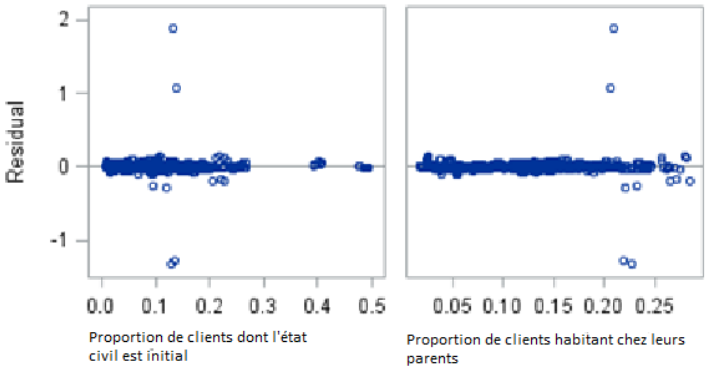


Résidus en fonction des quantiles de la loi normale



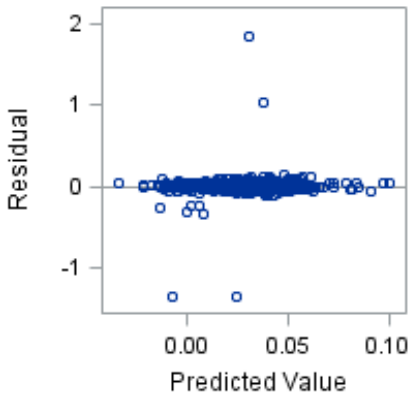
Résidus en fonction des variables dépendantes



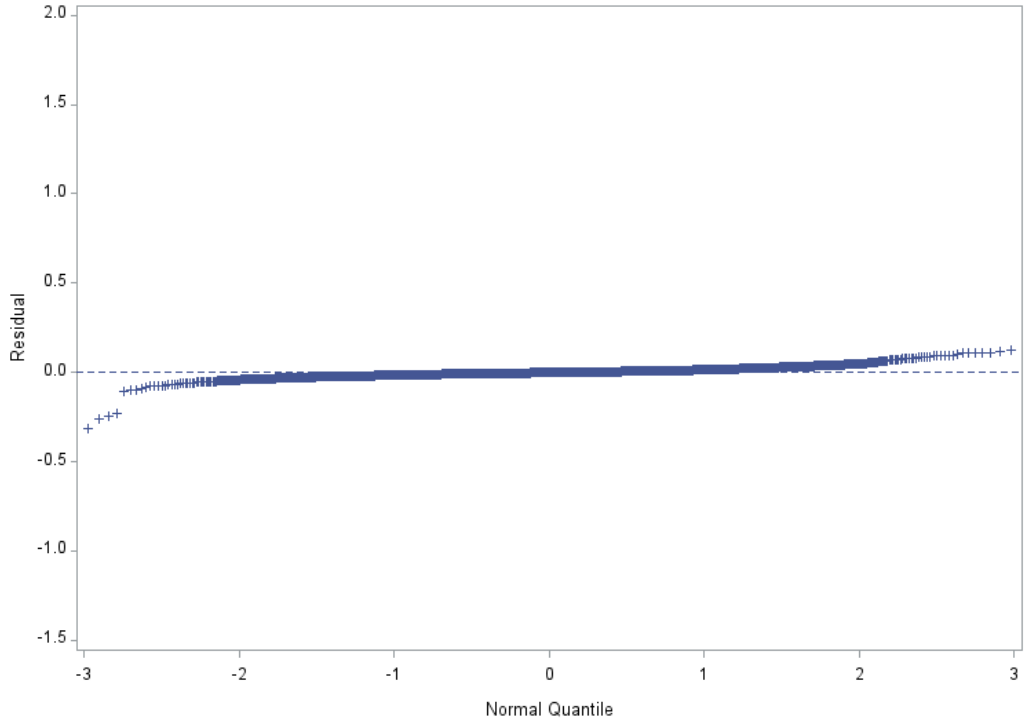

Residual
2
1
0
-1
0.0
0.1
0.2
0.3
0.4
0.5
Proportion de clients dont l'état civil est initial
0.05
0.10
0.15
0.20
0.25
Proportion de clients habitant chez leurs parents

Validation du modèle 2

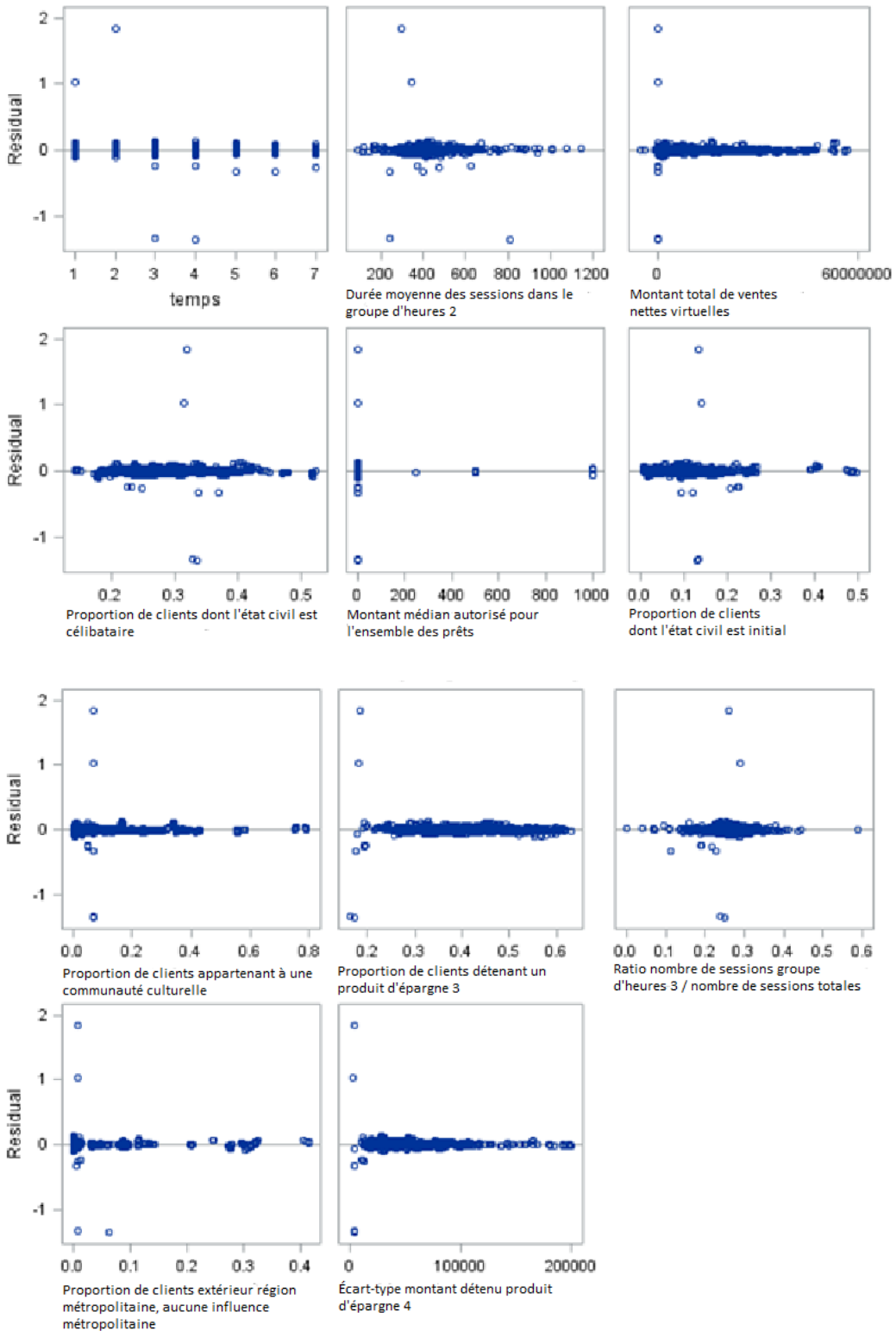
Résidus en fonction des valeurs prédites



Résidus en fonction des quantiles de la loi normale



Résidus en fonction des variables dépendantes



Bibliographie

Azmi, M., S. Araghinejad et M. Kholghi (2010). « MULTI MODEL DATA FUSION FOR HYDROLOGICAL FORECASTING USING K-NEAREST NEIGHBOUR METHOD* », *Iranian Journal of Science and Technology*, vol. 34, no B1, p. 81-92.

Baliga, Ganesh, Sanjay Jain et Arun Sharma (1997). « Learning from Multiple Sources of Inaccurate Data », *SIAM Journal on Computing*, vol. 26, no 4, p. 961-990.

Baud, Nicolas, Antoine Frachot et Thierry Roncalli (2002). « Internal Data, External Data and Consortium Data-How to Mix Them for Measuring Operational Risk », *External Data and Consortium Data-How to Mix Them for Measuring Operational Risk (June 1, 2002)*.

Bolancé, Catalina, Montserrat Guillén, Jim Gustafsson et Jens Perch Nielsen (2013). « Adding prior knowledge to quantitative operational risk models », *The Journal of Operational Risk*, vol. 8, no 1, p. 17-32.

Bradley, Cathy J., Lynne Penberthy, Kelly J. Devers et Debra J. Holden (2010). « Health Services Research and Data Linkages: Issues, Methods, and Directions for the Future », *Health Services Research*, vol. 45, no 5p2, p. 1468-1488.

Dahen, Hela et Georges Dionne (2010). « Scaling models for the severity and frequency of external operational loss data », *Journal of Banking & Finance*, vol. 34, no 7, p. 1484-1496.

Fernandez, G. C. (2007). « Model selection in PROC MIXED—a user-friendly SAS macro application », communication présentée au *SAS Global forum*, Orlando, Fla, USA.

Gamache, Philippe, Robert Pampalon et Denis Hamel (2010). « Guide méthodologique:«L'indice de défavorisation matérielle et sociale: en bref» », *Québec, Institut national de santé publique du Québec, septembre*.

Gustafsson, Jim et Jens Perch Nielsen (2008). « A mixing model for operational risk », *Available at SSRN 1090863*.

Hamel, Denis, Robert Pampalon et Philippe Gamache (2009). *Guide d'utilisation du programme d'assignation de l'indice canadien de défavorisation matérielle et sociale, année 2006*, Institut national de santé publique du Québec.

Heimbigner, Dennis et Dennis McLeod (1985). « A federated architecture for information management », *ACM Trans. Inf. Syst.*, vol. 3, no 3, p. 253-278.

Higgins, Aparna, Timothy Zeddies et Steven D. Pearson (2011). « Measuring The Performance Of Individual Physicians By Collecting Data From Multiple Health Plans: The Results Of A Two-State Test », *Health Affairs*, vol. 30, no 4, p. 673-681.

Horton, Nicholas J., Richard Saitz, Nan M. Laird et Jeffrey H. Samet (2002). « A Method for Modeling Utilization Data from Multiple Sources: Application in a Study of

Linkage to Primary Care », *Health Services & Outcomes Research Methodology*, vol. 3, no 3-4, p. 211-223.

Keselman, Harvey J., James Algina, Rhonda K. Kowalchuk, Russell D. Wolfinger et Matthew J. Gurka (2006). « Selecting the Best Linear Mixed Model under Reml, "The American Statistician," 60, 19-26: Comment by Keselman, Algina, Kowalchuk, and Wolfinger and Response », *The American Statistician*, vol. 60, no 2, p. 210-211.

Larocque, Denis (Automne 2014). *Analyse de données longitudinales et de survie - Recueil deuxième partie du cours*, Service de l'enseignement des méthodes quantitatives de gestion, HEC Montréal.

Larocque, Denis (hiver 2013). *Notes de cours : Analyse multidimensionnelle appliquée (6-602-07)*, Service de l'enseignement des méthodes quantitatives de gestion, HEC Montréal.

Molitor, Nuoo-Ting, Nicky Best, Chris Jackson et Sylvia Richardson (2009). « Using Bayesian graphical models to model biases in observational studies and to combine multiple sources of data: application to low birth weight and water disinfection by-products », *Journal of the Royal Statistical Society. Series A, Statistics in Society*, vol. 172, no 3, p. 615-637.

Müller, Samuel, JL Scealy et AH Welsh (2013). « Model selection in linear mixed models », *Statistical Science*, vol. 28, no 2, p. 135-167.

Ngo, L et R Brand (2002). « Model Selection in Linear Mixed Effects Models Using SAS Proc Mixed », communication présentée au *SAS Users. Group International (22) San Diego*, March 16-19, 1997, California.

Pampalon, Robert, Denis Hamel, Philippe Gamache, Mathieu D Philibert, Guy Raymond et André Simpson (2012). « Un indice régional de défavorisation matérielle et sociale pour la santé publique au Québec et au Canada », *L'utilisation contemporaine des indicateurs socioéconomiques régionaux*, vol. 103, p. 17.

Reese, C. Shane, Alyson G. Wilson, Jiqiang Guo, Michael S. Hamada et Valen E. Johnson (2011). « A Bayesian Model for Integrating Multiple Sources of Lifetime Information in System-Reliability Assessments », *Journal of Quality Technology*, vol. 43, no 2, p. 127-141.

Tufféry, Stéphane (2010). « Titre: Data mining et statistique décisionnelle. L'intelligence des données. Editeur: Editions Technip Paris, 2010 3e éd. Format: 17 cm x 24 cm, 725 p. Bibliogr. p. 689-699, Index ».

Wolfinger, Russell D. (1996). « Heterogeneous Variance: Covariance Structures for Repeated Measures », *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 1, no 2, p. 205-230.

Zeng, Jihong et Chengxiu Tang (2008). « Overview of Research and Practices in Information Sharing for Enterprise Resource Planning », *The Journal of Applied Business and Economics*, vol. 8, no 4, p. 34-43.

Pages web consultées

« Indice de défavorisation, Canada, 2006 ». 2008 (1er août). Institut national de santé publique du Québec [en ligne]. <http://www2.inspq.qc.ca/santescope/indicedefavo.asp?NoIndD=9>. Consulté le 29 novembre 2014.

« Le Code Hash ». 2007 (1er mars). SAS [en ligne]. <http://www.sas.com/offices/europe/france/services/support/articles/US200703_a1.html>. Consulté le 29 novembre 2014.

« Premiers pas en régression linéaire avec SAS ». 2006. Revue MODULAD, Numéro 35 [en ligne]. <https://www.rocq.inria.fr/axis/modulad/numero-35/Tutoriel-confais-35/confais-35.pdf>. Consulté le 29 novembre 2014.

« ODS Table Names ». 2014. SAS/STAT® 9.2 User's Guide, Second Edition [en ligne]. <http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_mixed_sect026.htm>. Consulté le 29 novembre 2014.

« The Google Geocoding API ». 2014 (19 novembre). Google Maps API Web Services [en ligne]. <https://developers.google.com/maps/documentation/geocoding/>. Consulté le 29 novembre 2014.

« The MIXED Procedure ». 2014. SAS/STAT® 9.2 User's Guide, Second Edition [en ligne]. <http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#mixed_toc.htm>. Consulté le 29 novembre 2014.

« The REG Procedure ». 2014. SAS/STAT® 9.2 User's Guide, Second Edition [en ligne]. <http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#reg_toc.htm>. Consulté le 29 novembre 2014.

« Using the Google Geocoding API in Excel ». 2012 (24 mai). Police Analyst [en ligne]. <http://policeanalyst.com/using-the-google-geocoding-api-in-excel/>. Consulté le 29 novembre 2014.

Autres produits consultés

Fichier des attributs géographiques, guide de référence, Recensement de 2011. Produit no 92-151-G au catalogue de Statistique Canada.

Fichier des attributs géographiques,Recensement de 2011. Produit no 92-151-X au catalogue de Statistique Canada.