IIROC OCRCVM
Investment Industry Regulatory Organization of Canada
Organisme canadien de réglementation du commerce des valeurs mobilières

Classification des groupes de négociation
Méthodologie et résultats

Examen et analyse des opérations
Organisme canadien de réglementation du commerce des valeurs mobilières

Le 9 septembre 2014

Remerciements:

Baiju Devani, directeur de l'analytique, OCRCVM
Ad Tayal, scientifique en chef des données, OCRCVM
Lisa Anderson, chef de l'équipe de l'analytique, OCRCVM
Dawei Zhou, analyste-programmeur technique principal, OCRCVM
Juan Gomez, analyste des opérations, OCRCVM
Graham W. Taylor[1,2]

[1] Consultant à OCRCVM
[2] École d'ingénieur, Université de Guelph, Guelph, ON, gwtaylor@uoguelph.ca

Table des matières

1. Introduction

Les marchés financiers et des titres de capitaux propres ont gagné en complexité au cours de la dernière décennie, sous l'influence des progrès technologiques, de la concurrence accrue et de l'arrivée de nouveaux participants sur le marché. Sur ce marché qui mise de plus en plus sur la technologie, il est devenu essentiel de définir différentes catégories de participants se distinguant par l'empreinte de leurs opérations et de se pencher sur les interactions et l'impact de ces groupes sur la structure fondamentale des marchés, notamment au chapitre de la qualité, de l'équité et de l'intégrité.

Nous vous présentons ici une méthodologie originale et rigoureuse permettant de classifier les différents groupes d'utilisateurs en s'appuyant sur un système d'apprentissage machine supervisé. Les travaux réalisés antérieurement par l'OCRCVM étaient axés sur une dimension donnée (ratio ordres/opérations, par exemple) ou sur un nombre déterminé de dimensions [1, 2]. La présente approche fait fond sur les travaux antérieurs de l'organisme, mais s'en écarte notablement à deux égards :

1. Nous appliquons un système d'apprentissage supervisé à l'aide de machines à vecteurs de support (SVM) pour la classification des groupes d'utilisateurs. L'utilisation de telles méthodes d'apprentissage automatisé est bien établie dans d'autres domaines et nos résultats démontrent leur efficacité dans le contexte présent.
2. Nous mettons à profit la richesse des données dont nous disposons pour constituer une série composée de plus de 200 éléments caractérisant le comportement de chaque utilisateur. Nos expérimentations indiquent que l'alliance d'une vaste série d'éléments et d'une méthodologie rigoureuse produit de meilleurs résultats de classification que l'utilisation d'une série d'éléments sélectionnés individuellement par des experts.

Dans le présent document, nous décrivons précisément notre méthodologie et présentons certains résultats permettant d'évaluer l'efficacité de l'algorithme de classification. Nous appliquons ensuite ce schéma de classification pour segmenter l'échantillon des utilisateurs sur une période d'étude et faire ressortir certaines interactions et mesures fondamentales pour chaque groupe.

2. Méthodologie

Dans la présente partie, nous décrivons la méthodologie utilisée pour déterminer le type d'activité associé à un identifiant (UserID) de négociation sur les marchés des titres de capitaux propres au Canada et nous analysons les résultats pour la période allant du 3 mars 2013 au 28 juin 2013. Il s'agit d'une période de stabilité sur le plan des modifications apportées aux politiques ou aux règles. De plus, durant cette période, l'OCRCVM disposait d'un groupe autodéclaré d'identifiants de détail qu'il pouvait utiliser dans le cadre de sa méthodologie.

Nous avons choisi d'utiliser le champ de l'identifiant pour classifier les flux de transactions. Quoiqu'imparfait, l'identifiant est en effet le champ le plus utile pour déterminer les types de flux de transactions de manière cohérente. Au départ, le marché attribuait aux courtiers un identifiant distinct pour chaque négociateur. Les pratiques de négociation gagnant en complexité et en automatisation, l'utilisation de l'identifiant s'est élargie et le flux des ordres passés par des identifiants individuels est aujourd'hui plus compliqué. Les exemples suivants illustrent cette complexité :

- Une entité donnée peut disposer de plusieurs identifiants que lui ont attribués différents marchés ou courtiers par l'intermédiaire desquels elle accède aux marchés.
- Plusieurs entités peuvent utiliser un identifiant donné pour certaines activités de négociation; par exemple, on peut utiliser un identifiant pour tous les flux d'ordres de détail exécutés.

Le tableau ci-dessous présente les grandes catégories (segments) des flux de négociation sur les marchés canadiens des titres de capitaux propres, selon le type de compte (données fournies à l'OCRCVM dans le cadre des exigences réglementaires) et la complexité des stratégies employées :

Tableau 1 : Ampleur et complexité de la stratégie, par type de compte

		Simple/modeste			**Complexe/vaste/intense**	
Type de compte	Client :	Détail	Spéculateur sur séance	Institution-nel	Fonds de couverture	Fournisseur de liquidité électronique
	Stocks :	Teneur de marché de lots irréguliers		Facilitation pour le compte de clients		Stratégies de courtage
	Non-client :	NC-Détail				
	Spécialiste :	Spécialiste du marché				
	Teneur de marché d'options :	Teneur de marché d'options				

Nous avons réparti les activités de négociation en quatre catégories :

1. Négociation à haute vitesse (NHV)
    - Comprend le groupe Fournisseur de liquidité électronique
    - Peut comprendre les groupes Fonds de couverture et Stratégies de courtage
2. Détail (DET)
    - Comprend les groupes Détail et NC-Détail
3. Spécialiste (SP)
    - Comprend les groupes Teneur de marché de lots irréguliers et Spécialiste du marché
4. Opérations vendeur/acheteur (V/A)

- Comprend les groupes Facilitation pour le compte de clients, Institutionnel et Stratégies de courtage
- Peut comprendre les groupes Spéculateur sur séance, Fonds de couverture et Teneur de marché d'options

L'idée est de trouver une règle automatique et objective permettant de ranger chaque identifiant dans l'une des quatre catégories. Nous nous penchons sur des méthodes d'apprentissage statistique pour classer chaque identifiant à l'aide d'une série de caractéristiques extraites mesurant les pratiques de négociation de l'identifiant. Nous qualifions manuellement une sous-série restreinte d'identifiants en nous appuyant sur nos connaissances de l'entité de négociation et nous utilisons cette série d'identifiants étiquetés pour l'apprentissage d'une règle inductive (apprentissage supervisé). Les méthodes d'apprentissage machine supervisé, telles que les machines à vecteurs de support et les forêts d'arbres décisionnels, se sont révélées très efficaces pour l'apprentissage de modèles efficients lorsque le nombre de caractéristiques est important relativement à la taille de l'échantillon, en l'absence (ou la quasi-absence) d'hypothèses sur des covariables [3, 4]. Nous appliquons une méthodologie expérimentale rigoureuse pour former les modèles et vérifier les résultats.

i. Extraction des caractéristiques

Nous mettons au point une série complète de plus de 200 éléments ou caractéristiques sur la base des activités de négociation quotidiennes de chaque identifiant. Les caractéristiques visent à mesurer le comportement global d'un identifiant sur une journée de négociation donnée, en couvrant les thèmes suivants. Pour chaque thème, nous avons donné des exemples des types de caractéristiques utilisées dans l'algorithme :

Tableau 2 : Thèmes et exemples de caractéristiques

Thème	**Exemples**
Opérations et ordres	• Valeur totale des opérations • Nombre d'ordres modifiés • Ratio ordres/opérations
Dynamique des stocks	• Pourcentage des opérations désignées comme dispensées de la mention à découvert[3] • Position nette

[3] Pour obtenir des renseignements complémentaires sur la désignation d'ordre « dispensé de la mention à découvert », veuillez consulter l'Avis 12-0300 de l'OCRCVM.

Mesures de la vitesse	• Rapidité de modification des ordres • Pourcentage d'ordres « simultanés »
Type de compte	• Pourcentage d'opérations associées à un type de compte donné (client) • Pourcentage d'opérations visant des lots irréguliers exécutées avec un spécialiste
Modalités	• Pourcentage d'ordres valable jusqu'à une date donnée • Pourcentage d'opérations ciblant les pôles opaques de liquidité
Titres négociés	• Nombre de titres individuels négociés • Pourcentage des opérations par marché
Applications et blocs	• Pourcentage du volume d'applications
Marché	• Pourcentage des opérations par marché
Gestion du coût des opérations	• Pourcentage d'opérations actives • Rabais nets

ii. Lissage des caractéristiques

Les identifiants peuvent changer de comportement d'un jour à l'autre. Nous commençons donc par calculer les caractéristiques quotidiennement pour chaque identifiant, puis nous calculons une moyenne quotidienne mobile sur un mois à partir des observations journalières. Les moyennes mobiles des caractéristiques sont utilisées par les modèles d'apprentissage machine pour représenter chaque identifiant. Cette méthode permet de lisser les variations du comportement quotidien de l'identifiant, tout en intégrant les variations à long terme de ses pratiques fondamentales.

iii. Traitement préliminaire

Certaines caractéristiques affichent de très hauts degrés d'asymétrie ou d'aplatissement, elles présentent des valeurs extrêmes (schéma de répartition à larges extrémités), ce qui peut dissimuler les tendances connexes. Par exemple, la Figure 1 illustre la répartition du ratio ordres/opérations et des rabais nets. Les valeurs (absolues) extrêmement élevées de la caractéristique masquent les variations des valeurs (absolues) plus basses.

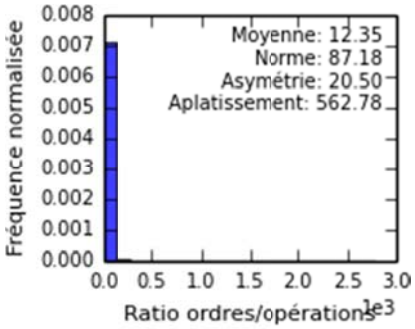
Pour améliorer la répartition d'une caractéristique et la visibilité des tendances de l'élément, nous transformons à l'aide de la formule suivante les caractéristiques dont le coefficient d'asymétrie est supérieur à 5 ou dont le coefficient d'aplatissement est supérieur à 20[4] :

$$\tilde{x} = sgn(x)\ln(1 + |x|),$$

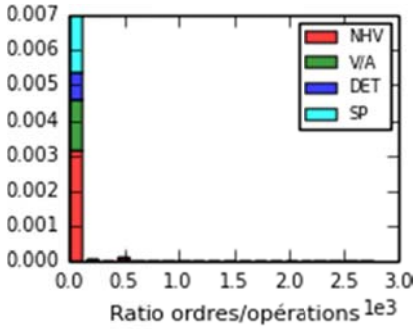
x étant la valeur d'origine de la caractéristique et $\tilde{x}$, sa valeur transformée. La formule prend en charge les valeurs positives aussi bien que négatives. La transformation logarithmique a pour effet d'élargir l'écart entre les valeurs (absolues) basses et de rapprocher les valeurs (absolues) élevées. Voir, par exemple, la Figure 2 plus bas, qui illustre la transformation logarithmique du ratio ordres/opérations et des rabais nets. La transformation logarithmique permet de mieux visualiser les tendances des caractéristiques. Ce traitement est particulièrement important pour les modèles linéaires qui sont incapables de s'adapter automatiquement à différentes représentations de caractéristiques. Le traitement préliminaire a également pour effet de rendre l'apprentissage plus stable numériquement.

À la dernière étape, nous normalisons chaque caractéristique pour obtenir une moyenne égale à zéro et un écart-type égal à un. Les valeurs manquantes reçoivent une valeur de zéro sur l'échelle normalisée, qui représente la valeur moyenne de la caractéristique.

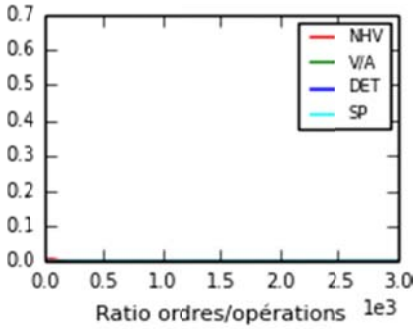
Figure 1 : Répartition du ratio ordres/opérations et des rabais nets – valeurs absolues



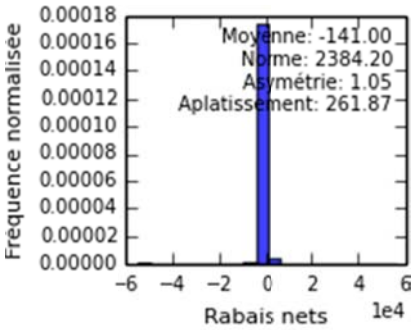
a) Histogramme des 2 009 échantillons



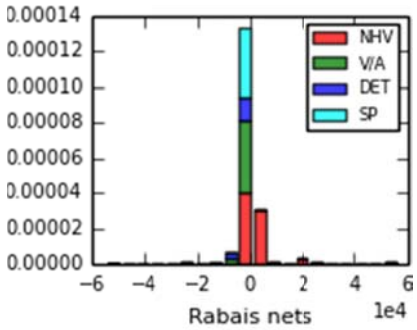
b) Histogramme des 105 échantillons étiquetés



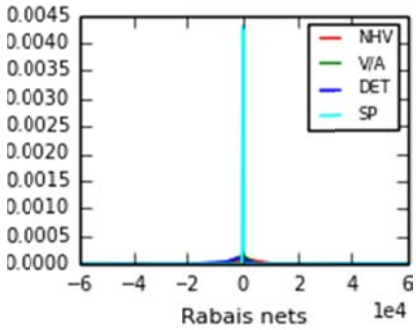
c) Densités conditionnelles des catégories, pour les échantillons étiquetés



a) Histogramme des 2 671 échantillons



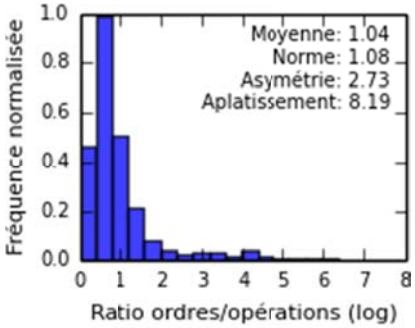
b) Histogramme des 113 échantillons étiquetés



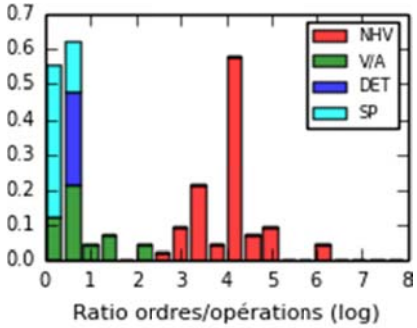
c) Densités conditionnelles des catégories, pour les échantillons étiquetés

[4] Dans le cadre de nos expérimentations, nous avons constaté que les résultats étaient insensibles aux valeurs limites précises choisies pour les coefficients d'asymétrie et d'aplatissement.

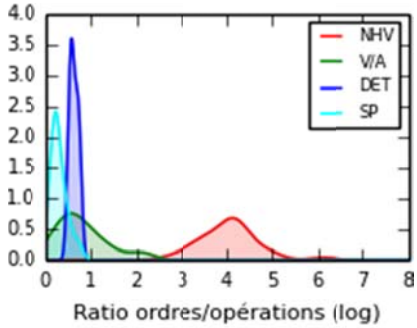
Figure 2 : Répartition du ratio ordres/opérations et des rabais nets – après transformation logarithmique



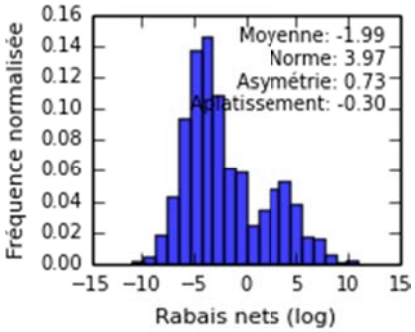
a) Histogramme des 2 009 échantillons



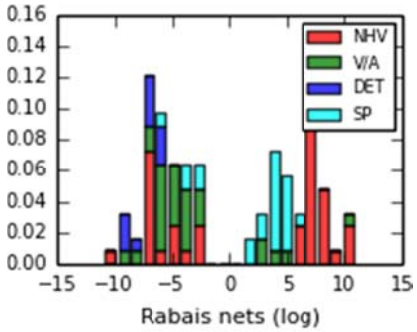
b) Histogramme des 105 échantillons étiquetés



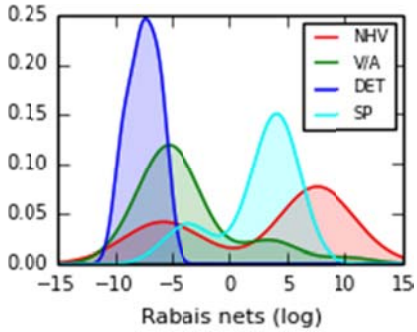
c) Densités conditionnelles des catégories, pour les échantillons étiquetés



a) Histogramme des 2 671 échantillons



b) Histogramme des 113 échantillons étiquetés



c) Densités conditionnelles des catégories, pour les échantillons étiquetés

iv. Série d'identifiants d'apprentissage étiquetés

Nous qualifions manuellement une série restreinte d'identifiants en nous appuyant sur notre connaissance de l'entité de négociation en date du 15 mai 2013. Le Tableau 3 présente la répartition des identifiants étiquetés[5].

Tableau 3 : Nombre d'identifiants étiquetés et non étiquetés au 15 mai 2013

	Catégorie de négociation	Nombre d'identifiants
Étiquetés	NHV	49
	V/A	29
	DET	11
	SP	24
Nombre total d'identifiants étiquetés		113
Nombre total d'identifiants non étiquetés		2 662
Pourcentage d'identifiants étiquetés		4,1 %

[5] Nous avons attribué les étiquettes en nous appuyant sur les connaissances préalablement acquises au sujet de l'entité dans le cours du travail réglementaire et analytique et sur les recherches effectuées à son sujet dans les sources de données publiques.

La série des identifiants étiquetés est utilisée pour former un algorithme d'apprentissage supervisé permettant de classer les identifiants dans l'une des quatre catégories.

v. Apprentissage et évaluation des modèles

Nous évaluons quatre types de modèles différents aux fins de la classification décrite.

1. Classifieur linéaire à machines à vecteurs de support (SVM) (« classifieur linéaire SVM ») [5]
2. Classifieur linéaire à machines à vecteurs de support (SVM) à pénalité à la norme L1 (« classifieur linéaire SVM à pénalité L1 ») [6]
3. Machines à vecteurs de support à base radiale (« SVM à base radiale ») [5]
4. Forêt d'arbres décisionnels [4]

Les trois méthodes utilisant des machines à vecteurs de support (1 à 3 de la liste ci-dessus) sont des classifieurs binaires. Elles sont adaptées au schéma multi-classe, en considérant chaque modèle de corrélation possible (six classifieurs binaires au total, rapprochant respectivement les catégories NHV à V/A, NHV à DET, NHV à SP, V/A à DET, V/A à SP et DET à SP). La prédiction finale repose sur un vote majoritaire. On parle de méthode « one-vs-one ». Les documents [7] et [8] traitent des avantages de l'utilisation de cette approche.

Pour évaluer la performance de chaque type de modèle, nous procédons comme suit. Nous commençons par diviser de façon aléatoire les données des identifiants étiquetés en deux séries, respectivement utilisées à des fins d'apprentissage et d'essais. La série d'apprentissage est composée d'un échantillon stratifié constitué de 80 % des données; la série d'essais est un échantillon stratifié qui comprend les 20 % restants. Nous utilisons un système de validation croisée à cinq volets pour la série d'apprentissage, afin d'ajuster les paramètres du modèle. Cette procédure évite la sélection d'un paramètre trop général pour les données d'apprentissage. Le modèle ajusté est ensuite évalué pour la série d'essais des 20 % laissés de côté. La procédure est répétée 20 fois pour obtenir des estimations de la moyenne et de l'écart-type, en vue de l'utilisation du modèle hors échantillon. Cette méthodologie fournit une évaluation précise de la performance hors échantillon de chaque type de modèle. Le Tableau 11 (annexe A) contient une liste des paramètres d'ajustement pris en considération pour chaque type de modèle.

Le Tableau 4 illustre la précision moyenne obtenue par chaque modèle pour les données hors échantillon. Le Tableau 5 et les suivants ci-dessous présentent la matrice de confusion moyenne pour chaque type de modèle. Les résultats montrent que tous les types de modèle permettent l'apprentissage efficace d'une règle de classification précise. Nous avons opté pour le classifieur linéaire SVM, qui obtient un niveau élevé de précision dans un espace d'hypothèse simple à fonction linéaire.

Pour le modèle définitif, en appliquant le classifieur SVM linéaire, nous utilisons des paramètres de pénalité distincts pour chaque problème de classification par paire. Ces paramètres sont sélectionnés à l'aide d'un système de validation croisée à huit volets visant la totalité des données. Les modèles

définitifs sont soumis à un nouvel apprentissage à partir de la totalité des données, en utilisant les paramètres optimaux retenus.

Tableau 4 : Précision moyenne et écart-type de chaque modèle en répétant la procédure 20 fois sur les données hors échantillon

Type de modèle	**Précision hors échantillon sur les exemples étiquetés**	**Écart-type**
Classifieur SVM linéaire	99,6	0,3
Classifieur linéaire SVM à pénalité L1	99,6	0,3
SVM à base radiale	98,5	0,6
Forêt d'arbres décisionnels	98,5	0,6

Tableau 5 : Matrice de confusion moyenne – Classifieur SVM linéaire (%)

		Prévision			
		NHV	**V/A**	**DET**	**SP**
Vrai	**NHV**	100	0	0	0
	V/A	0	98,3	1,7	0
	DET	0	0	100	0
	SP	0	0	0	100

Tableau 6 : Matrice de confusion moyenne – Classifieur linéaire SVM à pénalité L1 (%)

		Prévision			
		NHV	**V/A**	**DET**	**SP**
Vrai	**NHV**	100	0	0	0
	V/A	0	98,3	1,7	0
	DET	0	0	100	0
	SP	0	0	0	100

Tableau 7 : Matrice de confusion moyenne – SVM à base radiale (%)

		Prévision			
		NHV	**V/A**	**DET**	**SP**
Vrai	**NHV**	100	0	0	0
	V/A	4,2	95,0	0,8	0
	DET	0	2,5	97,5	0
	SP	0	0	0	100

Tableau 8 : Matrice de confusion moyenne – Forêt d'arbres décisionnels (%)

		Prévision			
		NHV	**V/A**	**DET**	**SP**
Vrai	**NHV**	100	0	0	0
	V/A	5,8	94,2	0	0
	DET	0	0	100	0
	SP	0	0	0	100

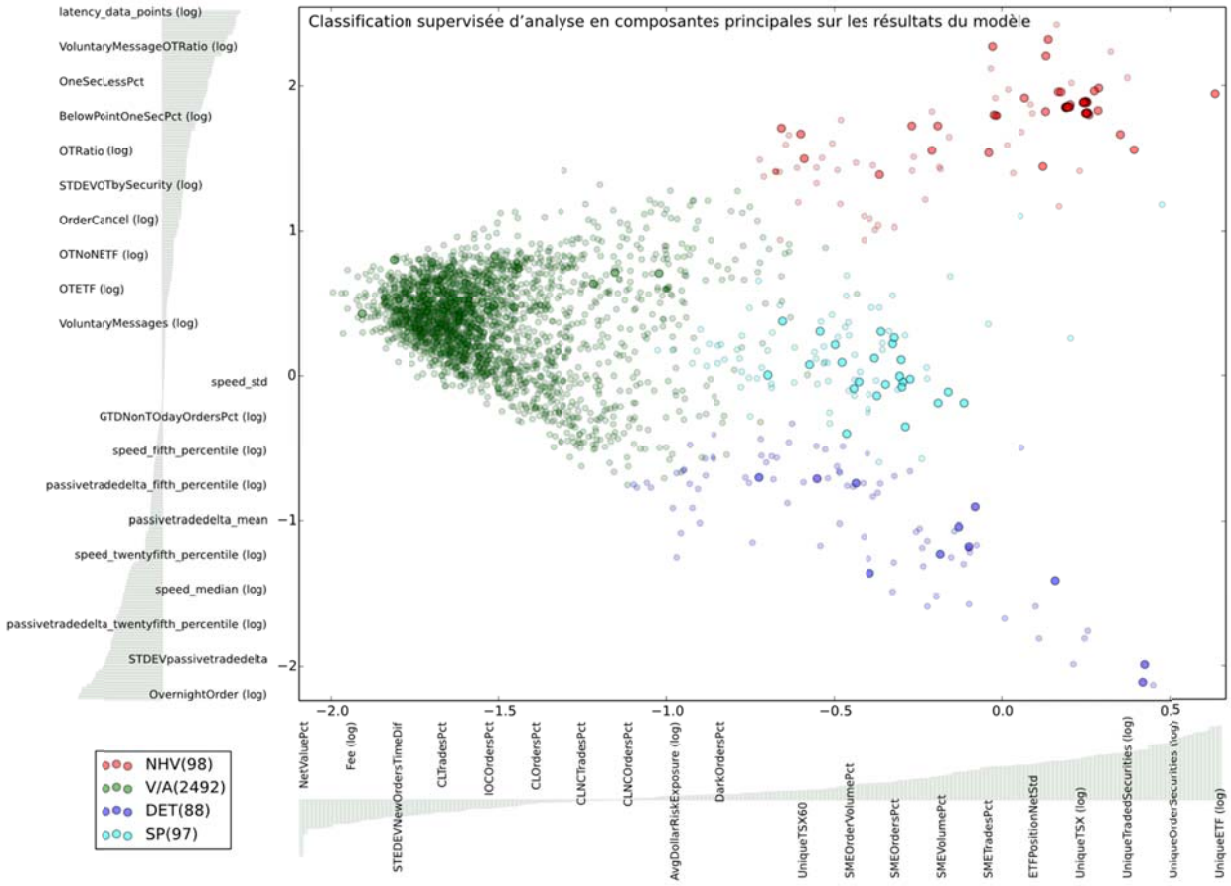
vi. Classification des identifiants

Nous utilisons le modèle définitif formé du classifieur SVM linéaire pour prévoir les étiquettes de tous les identifiants pour chaque mois de la période visée par l'étude. Le Tableau 9 illustre le nombre d'identifiants classés dans chaque catégorie de négociation au 15 mai 2013. La taille relative de chaque groupe cadre avec les résultats de notre série étiquetée et demeure stable d'un mois sur l'autre.
La Figure 3 projette les identifiants étiquetés sur un espace bidimensionnel à l'aide de l'algorithme de classification supervisée d'analyse en composantes principales afin de réduire la dimension et montre que le résultat du classifieur produit des groupes bien distincts d'identifiants.

Tableau 9 : Répartition des identifiants prévus au 15 mai 2013

Catégorie de négociation	**Nombre d'identifiants**
NHV	98
V/A	2 492
DET	88
SP	97
Total	2 775

Figure 3 : Visualisation des identifiants le 15 mai 2013, à l'aide de la méthode de classification supervisée d'analyse en composantes principales [9]



La méthode de classification supervisée d'analyse en composantes principales est principalement utilisée à des fins de visualisation; elle projette les catégories prévues par le classifieur SVM linéaire sur un espace bidimensionnel. La pondération des dimensions utilisées relativement à chaque composante est illustrée à l'aide de barres vertes.

vii. Stabilité de la classification au fil du temps

Aux fins de la présente évaluation et pour calculer les mesures des annexes B à F, nous avons effectué des prévisions pour chaque mois de la période visée par l'étude en nous basant sur la série de caractéristiques moyennes représentant l'activité mensuelle. Ce lissage des caractéristiques est nécessaire pour lisser la variabilité à court terme des classifications. Nous prévoyons néanmoins une certaine variabilité de la classification, sachant que de nouveaux identifiants sont créés, alors que d'autres disparaissent, et que certains identifiants pourraient représenter plusieurs clients et stratégies. La variabilité fait ici référence au changement de catégorie prévue d'un identifiant durant la période visée par l'étude.

Au total, 3 436 identifiants uniques ont été segmentés durant la période de quatre mois. La majorité des identifiants (98,2 %) se sont maintenus dans le même segment chaque mois. Les 62 autres identifiants (1,8 %) ont été classés dans plus d'un segment durant la période de quatre mois. L'incidence de ces identifiants à fourchette ambiguë représente 2 à 2,5 % du volume mensuel. Compte tenu de la stabilité générale de notre classification, tant au chapitre de l'univers des identifiants qu'à celui du volume qu'ils représentent, nous avons autorisé un identifiant à conserver l'étiquette prévue chaque mois.

3. Caractérisation des segments d'utilisateurs

Dans cette section, nous présentons certaines statistiques agrégées et interactions de l'univers de tous les identifiants classifiés à l'aide du processus susmentionné. Ces statistiques sont présentées en détail aux annexes B à E. La Figure 4 illustre les principales statistiques de négociation de l'ensemble des segments d'identifiants. Comme prévu, le groupe NHV est à l'origine de la plus grande proportion de messages d'ordres (91 %). Sur le plan du volume, la contribution du segment NHV est de l'ordre de 17 %. La Figure 6 représente le pourcentage de volume passif par segment d'utilisateur et montre que le groupe NHV est essentiellement passif (à 70 %). Cette constatation cadre avec nos attentes fondées sur la structure du marché canadien. Les annexes C et E illustrent les interactions parmi les groupes. Le groupe NHV, qui sert de contrepartie à 29 % de l'ensemble du volume des opérations, tend à être passif, indépendamment de sa contrepartie.

Nous quantifions également les activités de négociation de chaque groupe sur le plan des coûts d'opération. Nous procédons en utilisant l'écart effectif pondéré en fonction du volume et les écarts réalisés sur tous les titres du TSX 60. Le mode de calcul de ces mesures est présenté à l'annexe F. Ces mesures sont par ailleurs examinées en détail dans les documents [10, 11]. L'écart effectif, calculé pour chaque opération, traduit le différentiel payé (actif) ou perçu (passif) par rapport au prix médian du titre. L'écart réalisé traduit le coût lié à la sélection adverse (impact sur le cours) pour un négociateur, en

tenant compte du cours médian cinq minutes après l'opération[6]. La Figure 8 illustre les valeurs moyennes de ces mesures pour chaque groupe durant la période visée par l'étude.

Nous constatons que le groupe NHV gagne un écart effectif de 1,71 point de base (pb) correspondant à son flux global d'ordres passifs. Le groupe Détail, en revanche, paye un écart effectif de 1,11 pb, alors que le groupe V/A s'acquitte d'un écart effectif plus mince, de 1,03 pb. Au niveau des écarts réalisés, qui peuvent être considérés comme des profits nets pour les fournisseurs de liquidité [11], nous constatons que le groupe NHV gagne un écart réalisé de 0,14 pb (soit moins que l'écart effectif), ce qui suggère que ce groupe est principalement non directionnel. En moyenne, le groupe NHV gagne 0,77 pb par voie de rabais nets, soit environ 5,5 fois le revenu potentiel associé à la simple fourniture de liquidités. Dans le groupe V/A, dans lequel les flux sont a priori plus informés, l'impact sur les cours est positif, alors qu'il est négatif dans tous les autres groupes. Dans le groupe V/A, l'impact sur le cours excède le coût de l'écart effectif et se traduit par le gain d'un écart réalisé de 0,46 pb. Encore une fois, ces mesures agrégées cadrent avec nos prévisions.

Soulignons que ces mesures agrégées servent à caractériser les groupes et montrent que notre méthodologie de classification produit des résultats concordant avec nos attentes. Cependant, en réalisant la présente étude, nous ne tentons pas de caractériser l'impact de ces groupes sur la qualité, l'intégrité ou l'efficience du marché.

4. Résumé et prochaines étapes

Nous avons présenté une méthode de segmentation des participants au marché à partir de l'identifiant. L'approche d'apprentissage supervisé de la segmentation que nous avons décrite fait fond sur les vastes données dont dispose l'OCRCVM et sur nos connaissances des participants au marché. Les statistiques sommaires qui décrivent les groupes classifiés cadrent avec nos attentes et confirment la validité de la méthodologie que nous utilisons pour déterminer différents groupes. Cette méthode représente une amélioration par rapport à nos approches antérieures et un véritable progrès en direction de notre objectif à long terme. Nous entendons donner suite à ces travaux de plusieurs façons.

Premièrement, nous comptons appliquer notre modèle sur une plus longue période et opérationnaliser son utilisation pour approfondir notre étude de l'impact de chaque groupe sur les principales mesures de qualité, d'efficience et d'intégrité du marché.

Deuxièmement, nous continuerons à chercher des moyens d'élargir notre approche afin de définir des sous-groupes au sein de chaque segment. Comme d'autres l'ont signalé, certains groupes sont hétérogènes (notamment le segment NHV). Pour brosser un tableau plus précis, en raffinant la granularité, nous nous pencherons sur des méthodes nous permettant d'incorporer des caractéristiques reflétant la dynamique des marchés et les pratiques intrajournalières.

[6] Cet élément va dans le sens de la déclaration de l'écart réalisé prévu par la règle 605 de la Securities and Exchange Commission (SEC) et les travaux universitaires [12, 10].

Enfin, nous continuerons d'améliorer la méthodologie décrite dans le présent document. Par exemple, dans le cadre de ce travail, nous avons utilisé un algorithme d'apprentissage supervisé dans lequel les identifiants d'apprentissage étaient étiquetés en fonction des domaines de spécialisation. Nous souhaiterions étudier des modes de collaboration plus étroits afin d'étiqueter cette série d'identifiants en fonction de domaines de spécialisation multiples. Par ailleurs, bien que nous ayons constaté une amélioration des taux de classification lorsque nous utilisions une vaste série plutôt qu'une série restreinte de caractéristiques, nous jugeons avantageux de réduire le nombre de caractéristiques (attributs) aux fins de la classification.

Notre ambition à long terme est d'aller au-delà de la classification des différentes catégories d'utilisateurs décrites ici, et de réussir à repérer les habitudes ou les stratégies de négociation qui distinguent les groupes les uns des autres et de comprendre l'impact de chaque groupe sur les marchés.

5. Remarques à propos des annexes

Le Tableau 10 fournit des précisions à propos des analyses réalisées dans chacune des annexes suivantes pour en faciliter l'interprétation.

Tableau 10 : Remarques à propos des annexes

Remarques :	**Annexe**				
	B	**C**	**D**	**E**	**G**
Le segment est attribué à partir d'une prédiction mensuelle unique pour chaque identifiant; cette prédiction repose sur la série de caractéristiques moyennes représentant l'activité mensuelle.	x	x	x	x	x
Les statistiques quotidiennes sont cumulées pour chaque segment, avant de calculer une moyenne quotidienne; le volume moyen et la valeur moyenne font référence au volume négocié et à la valeur négociée	x	x	x	x	
L'étude couvre les activités de négociation survenues sur tous les marchés à l'égard de tous les titres cotés de 0 h à 24 h	x	x	x	x	
La catégorie AAPP décrit les opérations qui sont soit actives-actives soit passives-passives; ce type de négociation comprend par exemple les applications, les opérations au dernier cours, les opérations au premier cours et les opérations exécutées sur Match Now			x	x	
Les moyennes pondérées en fonction du volume ont été calculées quotidiennement pour chaque segment d'identifiant, avant d'en calculer une moyenne quotidienne					x
L'analyse a été limitée aux activités de négociation visant les titres du TSX 60 (sur tous les marchés) réalisées de 9 h 30 à 16 h					x
Seules les opérations comportant un côté actif et un côté passif ont été prises en compte; les opérations AAPP (voir plus haut) ont été exclues					x
Le volume, la valeur et le nombre d'opérations sont comptabilisés deux fois, dans le sens où l'acheteur et le vendeur comptabilisent chacun l'opération	x		x		x
Le volume, la valeur et le nombre d'opérations sont comptabilisés une seule fois, chaque opération n'est comptabilisée qu'une fois		x		x	

6. Annexe A

Le Tableau 11 contient une liste des paramètres d'ajustement pris en considération pour chaque type de modèle.

Tableau 11 : Paramètres d'ajustement pour chaque type de modèle

Type de modèle	**Paramètre**	**Série de valeur(s)**
Classifieur SVM linéaire	Pénalité d'erreur de classification (C)	2^[-15,-14,…,14,15]
Classifieur linéaire SVM à pénalité L1	Pénalité d'erreur de classification (C)	2^[-15,-14,…,14,15]
SVM à base radiale	Pénalité d'erreur de classification (C) Largeur du noyau à base radiale (σ^2)	2^[-15,-14,…,14,15] 2^[-10,-9,…,9,10]
Forêt d'arbres décisionnels	Nombre d'arbres	500

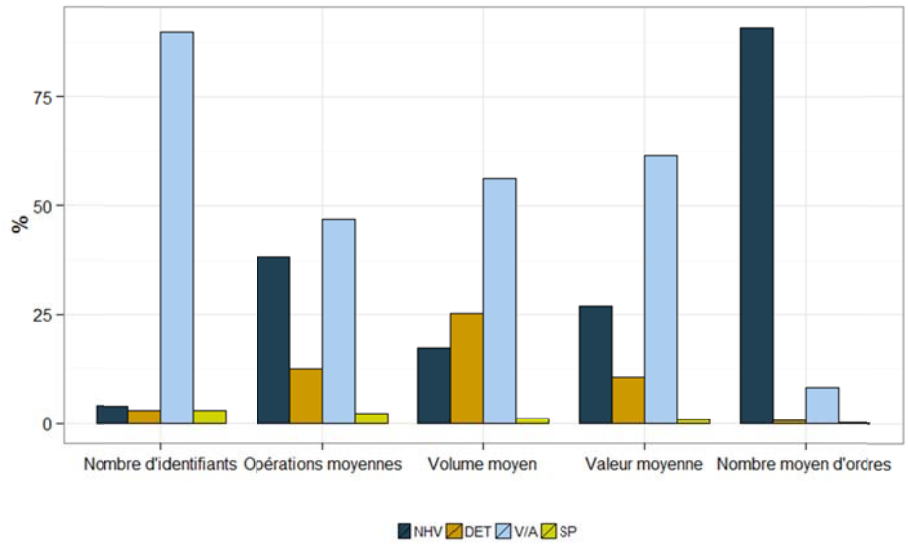
7. Annexe B : Statistiques sommaires par segment d'identifiants

Le Tableau 12 et la Figure 4 illustrent les statistiques sommaires par segment d'identifiants.

Tableau 12 : Statistiques sommaires moyennes quotidiennes – Pourcentage par segment d'identifiants

Segment d'identifiants	Nombre d'identifiants	Volume moyen	Valeur moyenne	Opérations moyennes	Nombre moyen d'ordres	Ratio ordres/ opérations moyen
NHV	4 %	17 %	27 %	38 %	91 %	55.4
DET	3 %	25 %	11 %	13 %	1 %	1.1
V/A	90 %	56 %	61 %	47 %	8 %	4.1
SP	3 %	1 %	1 %	2 %	0 %	3.2

Figure 4 : Statistiques sommaires moyennes quotidiennes – Pourcentage par segment d'identifiants



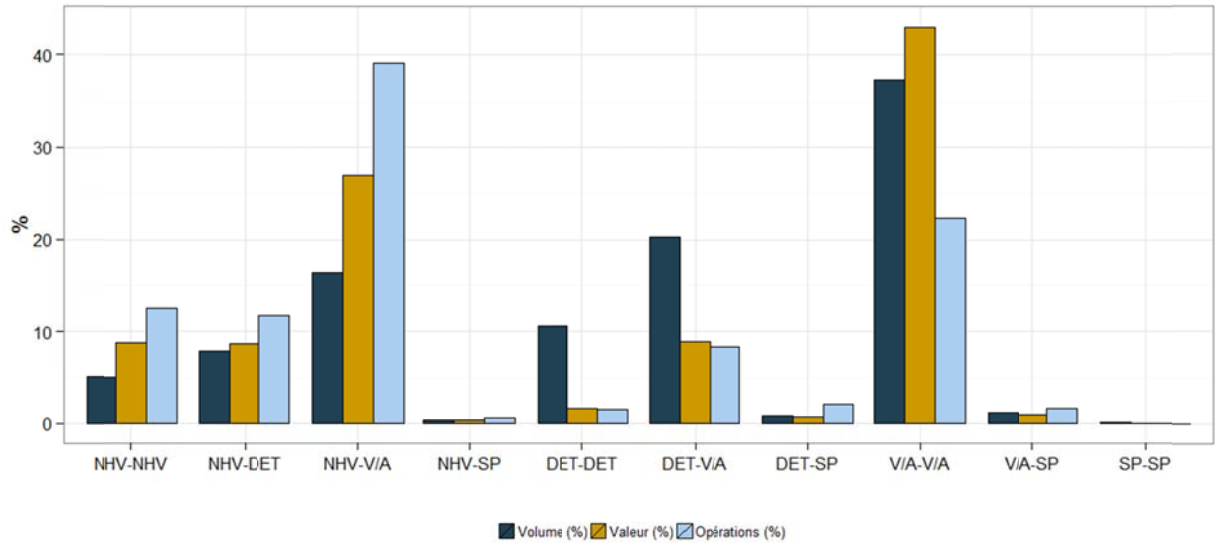
8. Annexe C : Statistiques de négociation par segment d'identifiants et contrepartie

Le Tableau 13 et la Figure 5 rapportent les statistiques de négociation par segment d'identifiants et contrepartie.

Tableau 13 : Moyennes quotidiennes du volume, de la valeur et du nombre d'opérations – Pourcentage par segment d'identifiants et contrepartie

Identifiant	Identifiant de la contrepartie	Volume moyen	Valeur moyenne	Opérations moyennes
NHV	NHV	5 %	9 %	12 %
NHV	DET	8 %	9 %	12 %
NHV	V/A	16 %	27 %	39 %
NHV	SP	0 %	0 %	1 %
DET	DET	11 %	2 %	2 %
DET	V/A	20 %	9 %	8 %
DET	SP	1 %	1 %	2 %
V/A	V/A	37 %	43 %	22 %
V/A	SP	1 %	1 %	2 %
SP	SP	0 %	0 %	0 %

Figure 5 : Moyennes quotidiennes du volume, de la valeur et du nombre d'opérations – Pourcentage par segment d'identifiants et contrepartie



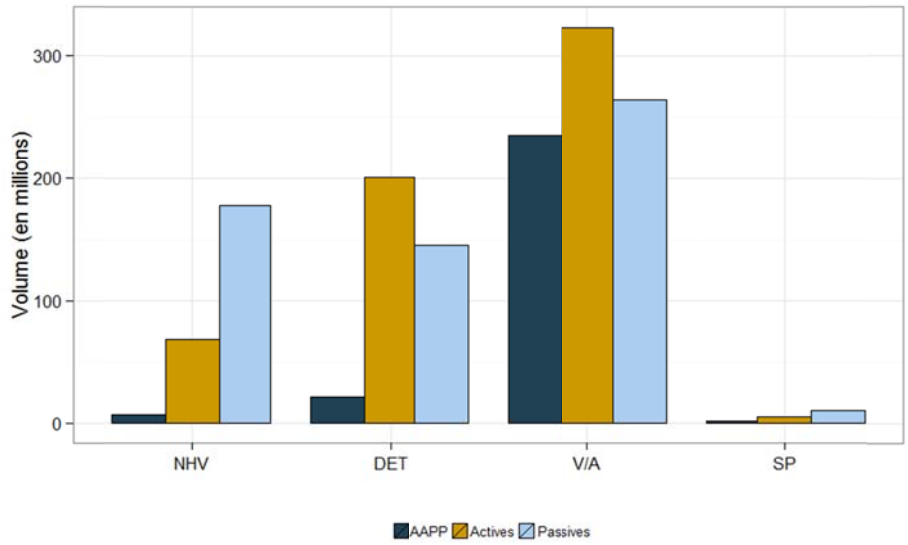
9. Annexe D : Volume des opérations actives et passives par segment d'identifiants

Le Tableau 14 et la Figure 6 illustrent le volume et la proportion d'opérations actives et passives par segment d'identifiants.

Tableau 14 : Pourcentage moyen quotidien d'opérations actives et passives – par segment d'identifiants

Segment d'identifiants	AAPP	Actives	Passives
NHV	3 %	27 %	70 %
DET	6 %	54 %	40 %
V/A	29 %[7]	39 %	32 %
SP	9 %	30 %	61 %

Figure 6 : Volume moyen quotidien d'opérations actives et passives – par segment d'identifiants



[7] La forte proportion d'opérations AAPP est attribuable aux applications intentionnelles, ce qui est prévu pour ce groupe.

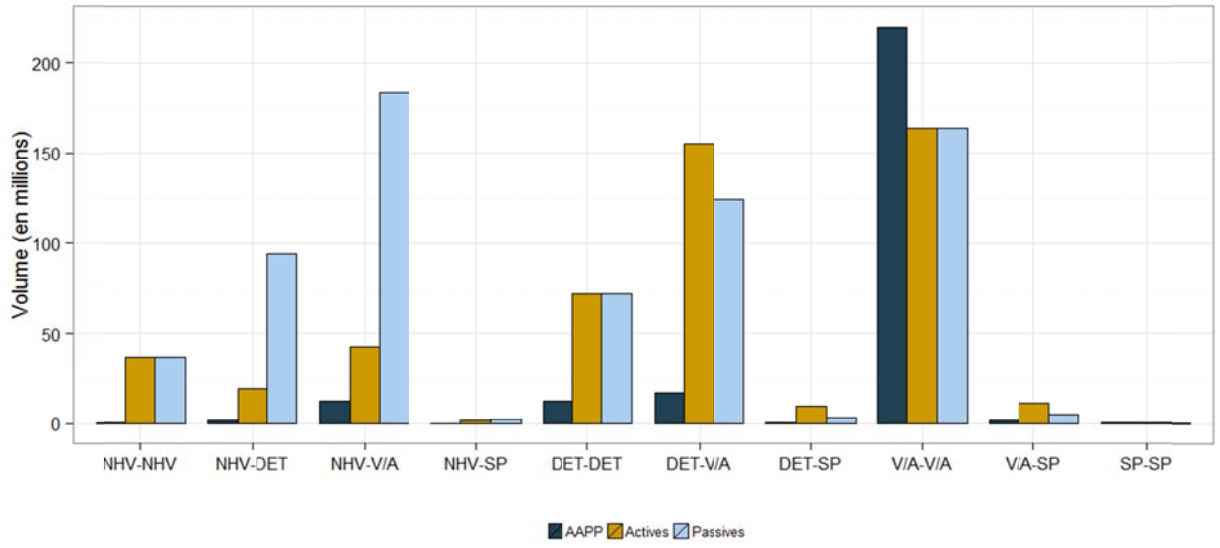
10. Annexe E : Volume des opérations actives et passives par segment d'identifiants et contrepartie

Le Tableau 15 et la Figure 7 illustrent le volume et la proportion d'opérations actives et passives par segment d'identifiants.

Tableau 15 : Pourcentage moyen quotidien d'opérations actives et passives – par segment d'identifiants et contrepartie

Identifiant	Identifiant de la contrepartie	AAPP par identifiant	Opérations actives par identifiant	Opérations passives par identifiant
NHV	NHV	0 %	50 %	50 %
NHV	DET	1 %	17 %	82 %
NHV	V/A	5 %	18 %	77 %
NHV	SP	2 %	37 %	61 %
DET	DET	8 %	46 %	46 %
DET	V/A	6 %	52 %	42 %
DET	SP	5 %	69 %	26 %
V/A	V/A	40 %	30 %	30 %
V/A	SP	8 %	64 %	28 %
SP	SP	36 %	32 %	32 %

Figure 7 : Volume absolu moyen quotidien d'opérations actives et passives – par segment d'identifiants et contrepartie



11. Annexe F : Coûts de négociation par segment d'identifiants - définitions

Semi-écart effectif (« EE »)

Le semi-écart effectif mesure la différence entre le cours de négociation (C_{it}) et la valeur actuelle du titre (i), représentée par le point médian de l'écart au moment de l'opération (V_{it}), divisée par le point médian de l'écart au moment de l'opération. La formule est la suivante, D_{it} étant une variable qui prend la valeur +1 pour l'acheteur et -1 pour le vendeur [10] :

$$EE_{it} = D_{it} * \frac{(C_{it} - V_{it})}{V_{it}}$$

Pour l'acheteur, l'écart effectif est positif si le cours de négociation est supérieur au point médian (par exemple, lorsqu'un ordre d'achat actif traverse l'écart), et il est négatif si le cours de négociation est inférieur au point médian (par exemple, lorsqu'un ordre de vente actif traverse l'écart). Pour le vendeur, l'écart effectif est positif si le cours de négociation est inférieur au point médian et négatif si le cours de négociation est supérieur au point médian.

Impact sur le cours (« IC »)

L'impact sur le cours mesure la différence entre la valeur future du titre, représentée par le point médian de l'écart cinq minutes après l'opération (V_{it+5}), et la valeur actuelle du titre, représentée par le point médian de l'écart au moment de l'opération (V_{it}), divisée par la valeur actuelle du titre. La formule utilisée est la suivante :

$$IC_{it} = D_{it} * \frac{(V_{it+5} - V_{it})}{V_{it}}$$

Pour l'acheteur, l'impact sur le cours est positif si le cours augmente après une opération, et négatif si le cours baisse. Pour le vendeur, l'impact sur le cours est positif si le cours baisse après une opération, et vice versa.

Semi-écart réalisé (« ER »)

Le semi-écart réalisé mesure la différence entre le cours de négociation (C_{it}) et la valeur future du titre, représentée par le point médian de l'écart cinq minutes après l'opération (V_{it+5}), divisée par la valeur actuelle du titre. La formule utilisée est la suivante :

$$ER_{it} = D_{it} * \frac{(C_{it} - V_{it+5})}{V_{it}}$$

Pour l'acheteur, l'écart réalisé est positif si le cours médian futur est inférieur au cours de négociation, et négatif s'il est supérieur. Pour le vendeur, l'écart réalisé est positif si le cours médian futur est supérieur au cours de négociation, et vice versa.

Les trois mesures sont reliées par la formule suivante :

$$ER = EE - IC$$

Frais du marché (frais et rabais) (« FM »)

Les frais du marché donnent une valeur de grandeur des frais payés ou des rabais obtenus à la valeur actuelle du titre (V_{it}) afin de pouvoir établir une comparaison avec les écarts réalisés et effectifs (ou intégrer cet élément aux écarts). La formule est la suivante, la variable R_{it} étant sélectionnée dans le tableau 16 ci-dessous selon le marché sur lequel a été exécutée l'opération, et selon que l'identifiant réalisait l'opération du côté actif ou du côté passif :

$$FM_{it} = \frac{-1 * (R_{it})}{V_{it}}$$

Le Tableau 16 synthétise des données disponibles au public à propos des barèmes de frais antérieurs et actuels. Cette synthèse est suffisante, car l'analyse a été restreinte aux opérations sur des titres du TSX 60 qui avaient un côté actif et un côté passif. Dans le tableau ci-dessous, les rabais correspondent aux données positives et les frais aux données négatives (publiées par les marchés). Selon le mode de calcul de l'écart effectif et de l'écart réalisé, les chiffres négatifs dénotent un profit et les chiffres positifs, un coût. Ce qui explique pourquoi, dans la formule de calcul des frais du marché, les rabais et frais sont multipliés par le facteur -1. Si les frais du marché sont négatifs, cela indique que le négociateur perçoit un rabais net; s'ils sont positifs, le négociateur paye des frais nets.

Tableau 16 : Barème des rabais et des frais estimés pour la période (en dollars)

Marché	Opération active	Opération passive
ALF	-0,0028	0,0025
CHX	-0,0029	0,0025
CNQ	-0,0025	0,002
CX2	0,001	-0,0014
ICX	-0,0015	-0,0015
LIQ	-0,01	-0,01
OMG	-0,0006	0
PTX	-0,0025	0,002
TCM	-0,001	-0,001
TMS	-0,0009	0,0005
TSX	-0,0035	0,0031
TSXV	-0,0035	0,0031

12. Annexe G : Coûts de négociation par segment d'identifiants – constatations

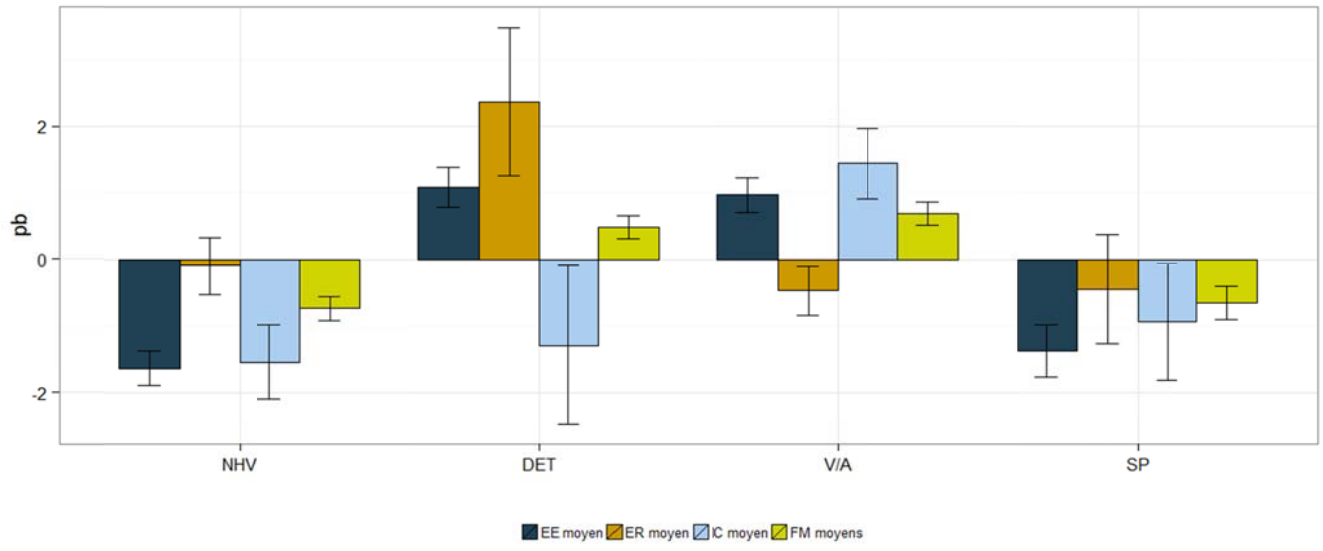
Le Tableau 17 et la Figure 8 illustrent les mesures des coûts de négociation attribués aux segments d'identifiants.

L'écart effectif, l'écart réalisé, l'impact sur le cours et les frais du marché ont été calculés pour l'acheteur et pour le vendeur de chaque opération, et attribués au segment d'identifiants pertinent. Les moyennes pondérées en fonction du volume ont été calculées quotidiennement pour chaque segment d'identifiants. Le Tableau 17 et la Figure 8 illustrent les valeurs moyennes des écarts, de l'impact sur le cours et des frais du marché. Les barres représentent l'écart-type par rapport à la moyenne.

Tableau 17 : Mesures du coût (en pb) par segment d'identifiants

Segment d'identifiants	Semi-écart effectif moyen	Semi-écart réalisé moyen	Moyenne de l'impact sur le cours	Moyenne des frais du marché
NHV	-1,71	-0,14	-1,57	-0,77
DET	1,11	2,43	-1,32	0,53
V/A	1,03	-0,46	1,49	0,72
SP	-1,36	-0,45	-0,91	-0,65

Figure 8 : Mesures du coût par segment d'identifiants



13. Bibliographie

[1] OCRCVM, « Étude des opérations ROOÉ ». Document de travail, 2012.

[2] OCRVM, « La qualité du marché dans un contexte en rapide évolution », Organisme canadien de réglementation du commerce des valeurs mobilières, 2013. Colloque de la CVMO et de l'OCRCVM sur la structure du marché - Le marché boursier canadien : Défis d'ordre structurel au milieu de changements rapides.

[3] V. N. Vapnik, *Statistical learning theory*. Wiley, 1re éd., 1998.

[4] L. Breiman, « Random forests », *Mach. Learn.*, vol. 45, no 1, p. 5–32, 2001.

[5] C. Cortes et V. Vapnik, « Support-vector networks », *Mach. Learn.*, vol. 20, no 3, p. 273–297, 1995.

[6] P. S. Bradley et O. L. Mangasarian, « Feature selection via concave minimization and support vector machines. », *ICML*, p. 82–90, Morgan Kaufmann, 1998.

[7] C.-W. Hsu et C.-J. Lin, « A comparison of methods for multiclass support vector machines », *Trans. Neur. Netw.*, vol. 13, no 2, p. 415–425, 2002.

[8] K. bo Duan et S. S. Keerthi, « Which is the best multiclass SVM method? An empirical study », *Proceedings of the Sixth International Workshop on Multiple Classifier Systems*, p. 278–285, 2005.

[9] E. Barshan, A. Ghodsi, Z. Azimifar et M. Zolghadri Jahromi, « Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds », *Pattern Recogn.*, vol. 44, no 7, p. 1357–1371, 2011.

[10] H. Bessembinder et K. Venkataraman, *Bid-Ask Spreads*. John Wiley & Sons, Ltd, 2010.

[11] K. Malinova, A. Park et R. Riordan, « Do retail traders suffer from high frequency traders? », publié sur le site SSRN à l'adresse http://ssrn.com/abstract=2183806 ou http://dx.doi.org/10.2139/ssrn.2183806, novembre 2013.

[12] SEC, *NMS Security Designation and Definitions*. 17 CFR Ch. II 242.600, 2013.