

Études des générateurs de déplacement à l'aide de données de cartes à puces

Martin Trépanier^{1,2}, professeur agrégé

Bruno Agard¹, professeur agrégé

Nicolas du Parc, étudiant au baccalauréat

1 – Centre interuniversitaire de recherche sur les réseaux d'entreprise, la logistique et le transport (CIRRELT)

2 – Groupe MADITUC en planification des transports urbains

École Polytechnique, département de mathématiques et de génie industriel

C.P. 6079, succursale Centre-Ville

Montréal, Québec

H3C3A7

Tél.: (514) 340-4711 #4911

Fax: (514) 340-4173

Courriel : [mtrepanier, bruno.agard, nicolas.du-parc]@polymtl.ca

Communication présentée au 44^e congrès de l'Association Québécoise du Transport et des Routes, Montréal, avril 2009

Études des générateurs de déplacement à l'aide de données de cartes à puces par Martin Trépanier, Bruno Agard et Nicolas Du Parc

Résumé

Les générateurs de déplacements sont des lieux qui causent et façonnent la mobilité urbaine. La connaissance des habitudes de la clientèle de ces grands pôles de destination est utile aux planificateurs du transport collectif urbain. Elle lui permet d'ajuster le service en fonction des volumes, du type d'usager et des heures de déplacements les plus prisées. Les grandes enquêtes origine-destination peuvent contribuer à l'étude des clientèles à un niveau stratégique, mais elles peuvent difficilement faire l'objet d'analyses sur une base continue car elles ne sont conduites qu'aux cinq ans.

Les systèmes de perception par carte à puce, tels que celui implanté à la Société de transport de l'Outaouais (STO) depuis 2001, permettent de caractériser les comportements de transport collectif associés aux cartes dans le plus strict anonymat des usagers. Dans cet article, nous présentons une méthode proposant l'analyse des générateurs de déplacements en utilisant comme unité analytique les arrêts d'autobus de la STO. Les habitudes de déplacements liés aux arrêts sont analysées à l'aide de méthodes de data mining en vue de les regrouper selon les comportements similaires et de déterminer à quels générateurs ils sont associés. À l'inverse, certains arrêts à proximité des grands générateurs du territoire sont identifiés et examinés pour valider l'approche.

Les expérimentations ont permis de faire ressortir les caractéristiques des catégories de générateurs les plus courants: hôpitaux (déplacements réguliers au cours de la journée), écoles (déplacements concentrés le matin et le soir), centre-ville (déplacements pendulaires) et centres commerciaux (soirs et week-end). En outre, une bonne régularité temporelle des comportements au cours de la semaine a pu être observée par l'examen de données longitudinales provenant du système de paiement par cartes à puces. Cependant, quelques variations locales sont perceptibles, comme l'augmentation de la clientèle des centres commerciaux le vendredi midi. Cette recherche dénote de nouvelles perspectives concernant l'utilisation potentielle de ces données pour l'ajustement du service.

Mots clé : système de paiement par cartes à puces, transport collectif, data mining, variabilité des habitudes de déplacements, générateurs de déplacements

1 Introduction

La Société de transport de l'Outaouais (STO) utilise un système de paiement de carte à puce depuis 2001. Chaque mois, la STO enregistre ainsi plus de 800 000 transactions dans sa base de données. L'historique des données liées aux techniques de fouille de données (data mining) et à l'exploitation de système d'information géographique permet de caractériser les habitudes de déplacement associées aux cartes (Morency et al. 2007).

En effet, chaque transaction validée avec une carte à puce est associée, entre autres, à un arrêt de bus, une ligne, un sens, une date et une heure de transaction. En recoupant ces données avec les emplacements géographiques des arrêts de bus et des générateurs de déplacement, il est possible de dresser un portrait global relativement fin des déplacements des usagers de la STO, tant en fonction du temps, que de l'espace, et ce malgré la variabilité de leurs habitudes. L'intérêt d'une telle approche est de caractériser au mieux le profil socio démographique des usagers et offre la possibilité de fournir ou de développer des services adaptés.

Cet article propose une méthodologie reposant sur des concepts de modélisation objet des données, de techniques de data mining tels que la classification (*clustering*), et l'utilisation d'un système d'information géographique (SIG). Après le rappel de certains fondements relatifs à la carte à puce, nous présentons dans la section « Systèmes d'information » les différentes étapes de traitement des données menant aux résultats de l'analyse, présentés à la section 0. La conclusion fait un retour sur les principaux constats et présente les différentes perspectives apportées par ce nouveau moyen de collecte de données.

2 Fondements

Les systèmes de cartes à puce sont en voie d'être généralisés pour la gestion du paiement dans les réseaux de transport collectif. Cette technologie permet une meilleure gestion des revenus, en plus de faciliter l'intégration tarifaire et les opérations de perception à bord des véhicules. L'utilité de tels systèmes est reconnue dans la littérature, malgré certaines limitations souvent occasionnées par les coûts importants d'implantation : les opérateurs décident de retirer certaines fonctions (nombre de titres, paiement électronique) plutôt de que de risquer de voir leur projet stoppé (Bonneau 2002). Ces systèmes soulèvent des questions quant à la confidentialité des données. Clarke (2001) rappelle que la carte à puce n'est qu'une des très nombreuses technologies

permettant de retracer l'identification et les habitudes d'une personne, parmi les systèmes bancaires, policiers, d'immatriculation, de péage routier et qu'en ce sens, des précautions similaires doivent être prises pour protéger les informations nominales.

Les avantages des systèmes de cartes à puce en transport collectif sont multiples (Bagchi et White 2005):

- un accès à un volume plus large de données sur les trajets individuels;
- la possibilité de relier les cartes à l'utilisateur;
- l'accès à des données continues couvrant de plus larges périodes de temps;
- la connaissance d'une grande proportion de sa clientèle.

En ce qui concerne les données de cartes à puces de la Société de transport de l'Outaouais, voici quelques dimensions qui ont été examinées dans le cadre de travaux des chercheurs de l'École Polytechnique :

- statistiques générales d'achalandage (Trépanier et al. 2004) ;
- détermination des arrêts de débarquement (Trépanier et al. 2007) ;
- variabilité dans les comportements des usagers mesurée à l'aide de techniques de *data mining* (Morency et al. 2007) ;
- techniques d'imputation des données erronées ou manquantes (Chu et Chapleau 2007) ;
- visualisation et modélisation des patrons de déplacements (Chapleau et Chu 2007, Chu et Chapleau 2008);
- comparaisons des données récoltées avec les résultats de l'enquête régionale origine-destination de 2005 (Trépanier et al. 2009).

Quant aux générateurs de déplacements, ils sont étudiés depuis longtemps par la communauté des planificateurs de transport, mais les données disponibles sont souvent agrégées ou très pointues, ne touchant que peu de sites. Les taux de génération de déplacements à différents sites typiques sont publiés régulièrement par *l'Institute of Transportation Engineers* américain (ITE 2003). Suite à différentes études des générateurs menées avec les données d'enquête ménage origine-destination dans la région de Montréal, les constats sont à l'effet que les générateurs de déplacements montrent des taux de déplacement très variables à l'intérieur d'une même région, et deux lieux proches peuvent attirer des clientèles très éparpillées et de distribution spatiale inégale (Trépanier et al. 2003).

Les données sont ensuite mises en forme pour leur utilisation dans Tanagra (Rakotomalala, 2005), un logiciel de *data mining*. Les techniques utilisées sont :

- le formatage des données (prétraitements, changements d'échelle, fusion de champs...)
- les filtres de données, qui permettent de déterminer des ensembles d'analyse;
- l'agrégation en groupes à l'aide de la méthode du K-moyen et du HAC (*Hierarchical Agglomerative Clustering*);
- la caractérisation des groupes obtenus (description, variabilité de l'appartenance aux groupes).

Ces techniques permettent de créer des groupes d'ALD montrant des comportements de déplacements communs. Ces déplacements sont liés ici aux montées enregistrées à ces ALD, pour lesquels des attributs limités sont disponibles : types de titre de transport, temporalité. De plus, il importe de vérifier l'appartenance d'un ALD à un groupe donné d'une journée à l'autre, ce qui permet de vérifier la stabilité longitudinale des comportements.

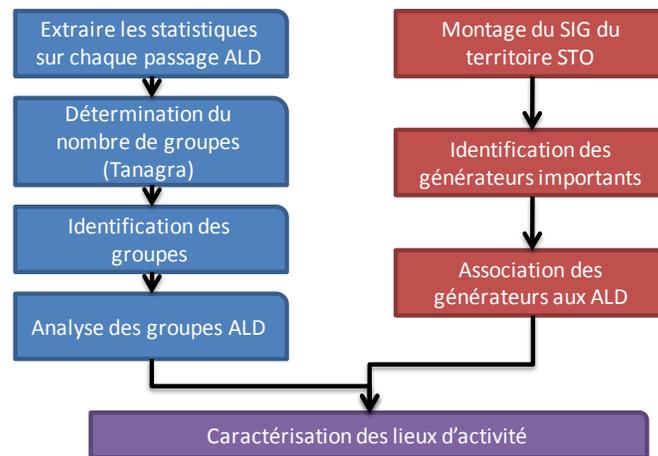


Figure 2: Méthodologie

La Figure 2 présente la méthodologie utilisée. D'un côté, les statistiques sur le nombre de montées à chaque passage sont extraites de la base de données, puis des groupes sont formés à l'aide du logiciel Tanagra. Les groupes sont ensuite identifiés puis leurs différentes caractéristiques sont analysées. En parallèle, les principaux générateurs de déplacements du territoire de la STO sont géolocalisés dans un système d'information géographique en vue de les associer par proximité aux ALD du réseau. Cette association permet enfin de caractériser ces lieux en fonction de l'usage du transport collectif.

4 Constats

Dans cette section, nous présentons les résultats des différentes expérimentations.

4.1 Caractérisation des groupes

Suite aux premières analyses, neuf groupes d'ALD ont été identifiés, chacun possédant un comportement d'usager propre. La Figure 3 présente la distribution des heures de montées pour chacun des groupes. Trois groupes sont associés à l'heure de pointe du matin (pic 6h, pic 7h et pic 8h), trois autres à l'heure de pointe du soir (15h, 16h, 17h). Le groupe « pic milieu » se caractérise par une forte activité entre 9h et 14h, tandis que le groupe « pic soir » touche la période entre 18h et 23h. Un dernier groupe, « pic régulier », échappe à la caractérisation par sa distribution régulière des montées.

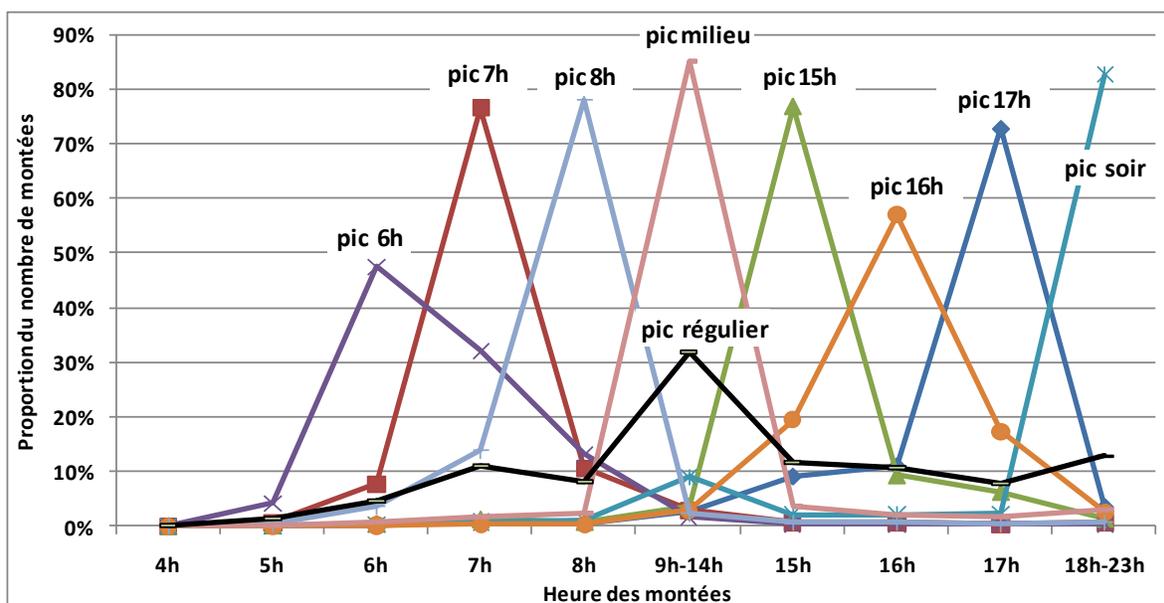


Figure 3: Distribution des heures de montée pour chacun des groupes

Chacun des ALD, pour chaque jour, est associé à un groupe. Le Tableau 1 présente les statistiques sur leur répartition dans les groupes et le nombre de montées associées. Le tableau permet également de connaître l'importance des classes de titres de transport, en fonction du nombre de montées. On y remarque que les groupes « pic 7h » et « pic régulier » sont les plus importants. De plus, les groupes « pic 6h » et « pic 16h » comptent une prépondérance d'adultes. Les étudiants sont concentrés dans les groupes centraux. La plus grande proportion d'aînés se retrouve quant à elle dans le groupe « pic milieu ».

Tableau 1: Statistiques sommaires sur les groupes, période du 1^{er} septembre au 31 décembre 2005

Groupe	ALD-jour	Montées
pic 6h	11,6%	10,8%
pic 7h	16,2%	10,5%
pic 8h	10,0%	5,3%
pic milieu	14,2%	6,5%
pic 15h	5,6%	5,9%
pic 16h	10,6%	15,7%
pic 17h	4,3%	2,2%
pic soir	6,5%	3,7%
pic régulier	21,2%	39,3%
	100%	100%
	(291829 ALD-jour)	(3172506 montées)

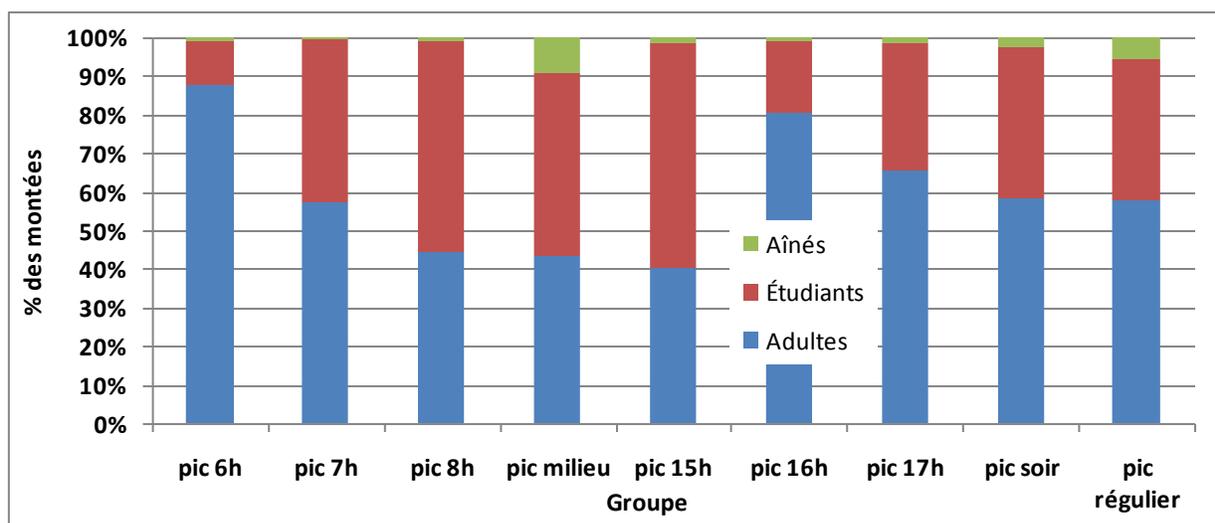


Figure 4: Catégories de titres impliquées dans les groupes

Chaque heure de la journée est associée à un pourcentage de montées appartenant à un groupe d'ALD donné. La Figure 4 présente ainsi la variabilité du pourcentage de monté par groupe en fonction des heures de la journée. On remarque qu'en dehors des périodes de pointes (7h à 8h et 15h à 17h), la majorité des usagers montent sur des ALD appartenant au groupe régulier. Pendant les heures de pointe, les usagers utilisent essentiellement des ALD dont le groupe à un pic pendant cette période. Ainsi, environ 60 % des montées ont lieu sur des ALD du groupe « pic 16h » à la période de pointe correspondante.

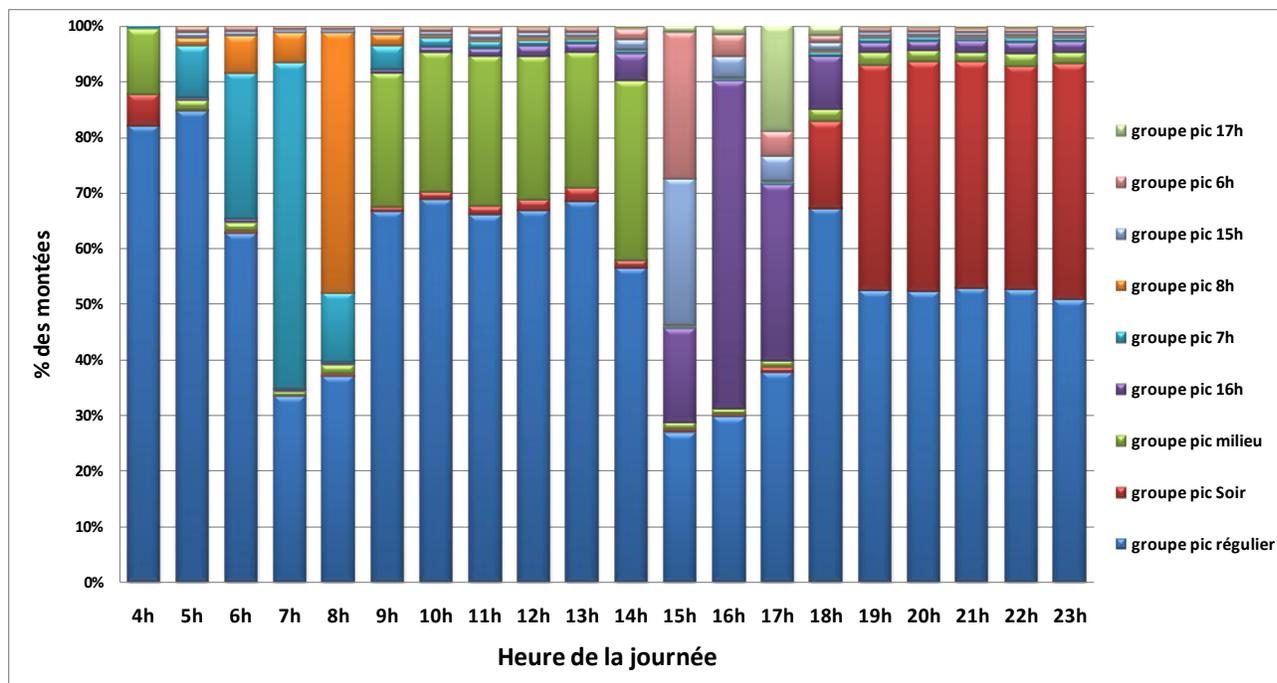


Figure 5: Pourcentage de montées par groupe et heure de la journée

4.2 Stabilité longitudinale

La répartition des comportements (groupes) n'est pas stable en fonction du temps. La Figure 6 présente la distribution de nombre de montées par groupe, par jour. On y remarque une variation sensible au cours des jours de semaine, ainsi que la modification des comportements qui est notoire durant le temps des Fêtes. Durant les week-ends, le groupe « pic milieu » prend une plus grande importance, les groupes de pointe étant rarement représentés.

Puisque un ALD peut changer de groupe comportemental d'appartenance au fil des jours, nous définissons la notion de groupe dominant, qui est celui comptant le plus grand nombre de montées à un ALD donné. Nous conservons aussi, pour chaque ALD, le nombre de groupes différents qui y est associé durant la période. La Figure 7 présente la répartition du nombre d'ALD selon ces deux paramètres. On y observe que lorsque le groupe dominant d'un ALD est « pic 6h », pic « 17h », pic « 16h » ou « pic 15h », il y a rarement plus de 5 groupes différents à cet ALD. Cela signifie que pour ces groupes, le comportement est très stable aux ALD. En contrepartie, les ALD où le groupe « pic régulier » est dominant peuvent compter jusqu'à 9 groupes différents, ce qui montre une grande variabilité.

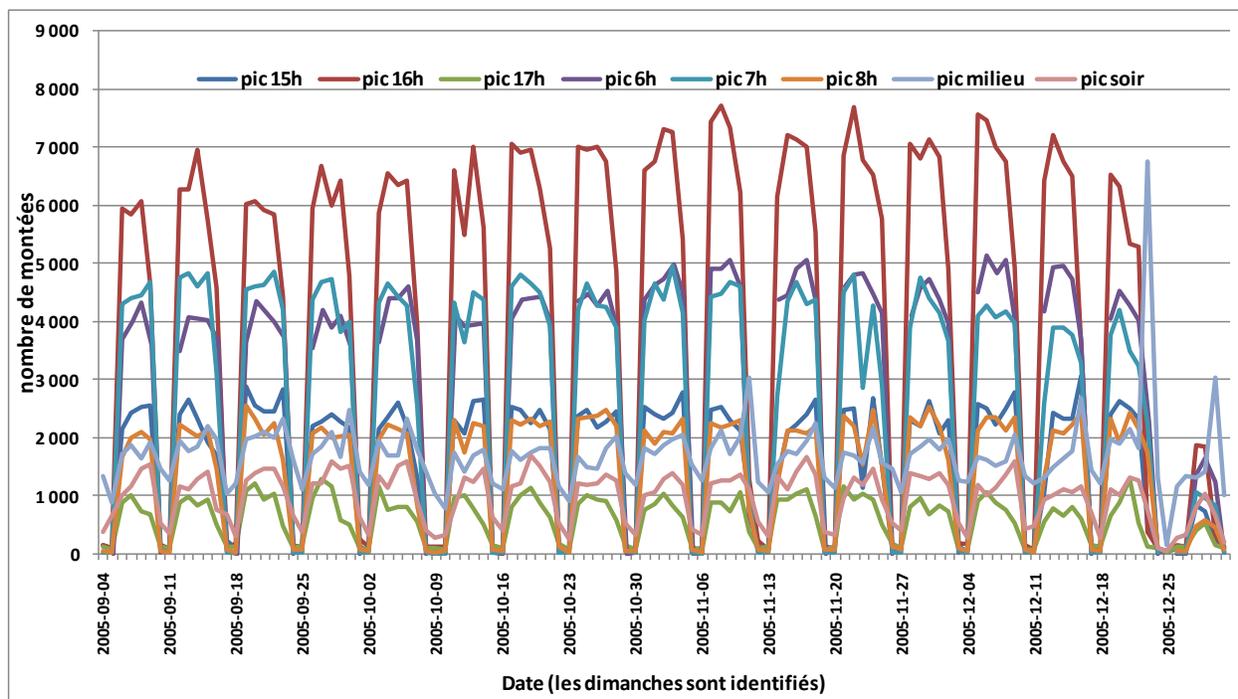


Figure 6: Distribution du nombre de montées par jour (groupe régulier omis)

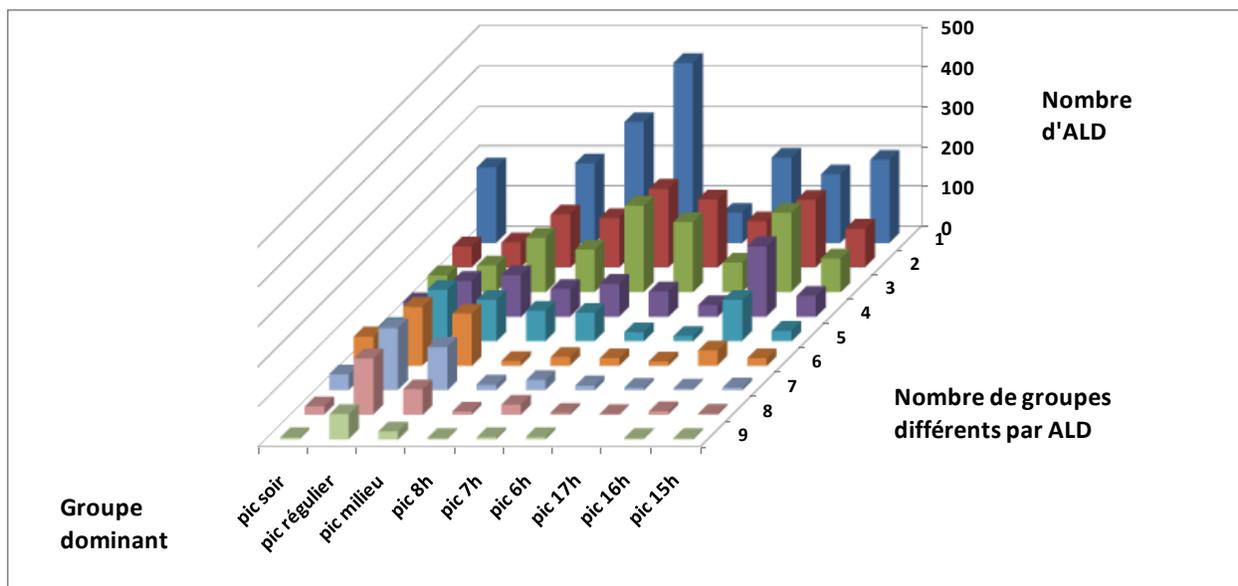


Figure 7: Répartition des ALD selon le nombre différents de groupes, par groupe dominant

4.3 Distribution spatiale

La distribution spatiale des ALD par groupe est variable en fonction des secteurs géographiques. La Figure 8 présente ainsi la distribution des ALD de chaque groupe. On constate plusieurs éléments

intéressants. Tout d'abord, plus on s'approche des zones commerciales et de l'hôpital, plus la proportion de montées appartenant au groupe « pic milieu » est importante. De même, plus on s'approche des secteurs résidentiels, plus la proportion de montées appartenant au groupe « pic 7h » est prédominante. Enfin, pour le secteur du casino, ce sont clairement les montées du groupe « pic soir » qui prédominent. Chaque groupe étant composé d'ALD, on obtient ainsi la distribution sectorielle des ALD en fonction de la prédominance d'un groupe. La prédominance d'un groupe en fonction de l'emplacement géographique d'un arrêt semble cohérente. En effet, il est logique d'avoir une prédominance du groupe « pic soir » sur les arrêts à proximité du casino, l'essentiel des activités de ce type de générateur de déplacement ayant lieu en soirée.

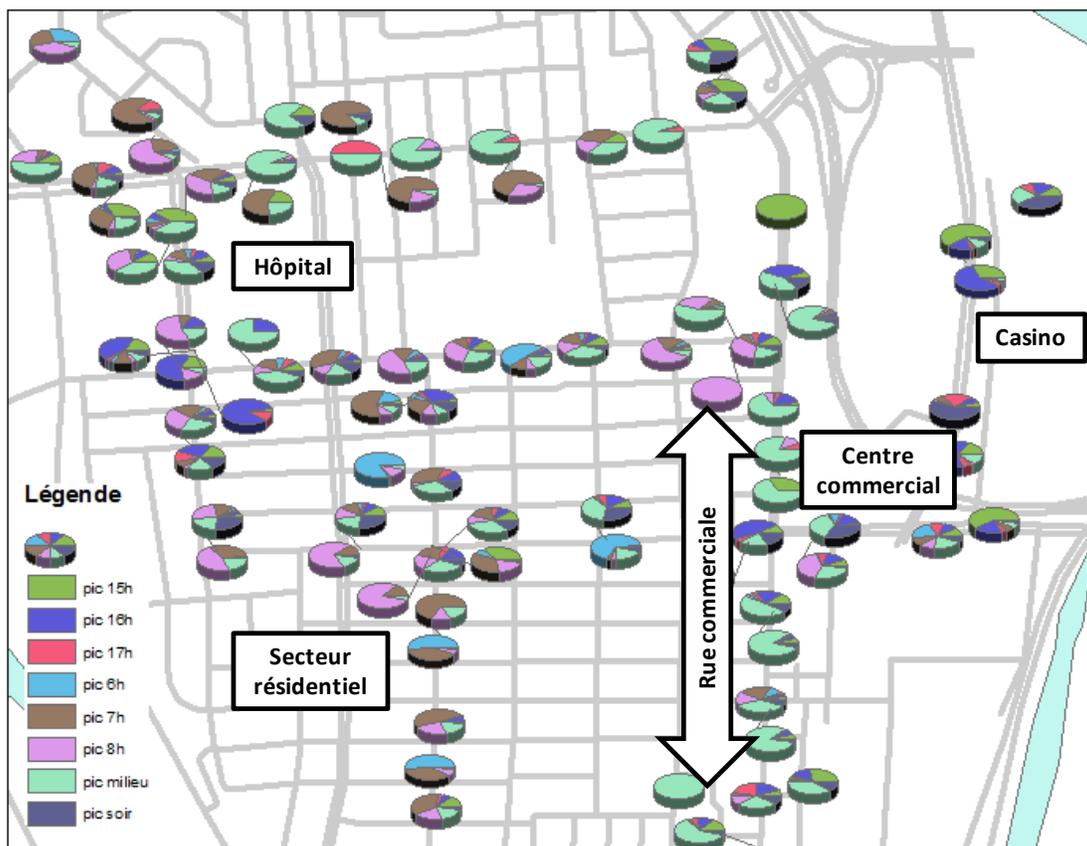


Figure 8: Carte des arrêts selon la répartition des montées par groupe (groupe régulier omis)

4.4 Analyse des générateurs

L'influence spatio-temporelle des générateurs de déplacement sur le réseau de transport en commun montre clairement des phénomènes de migration pendulaire. Ainsi, la relation de proximité entre les ALD et les générateurs de déplacements (Figure 9) donnent plusieurs

enseignements. Le centre-ville et la zone industrielle sont la source de la majorité des montées correspondant au pic de 16 heures, ce qui s'explique par la population quittant le travail. Si l'on s'intéresse aux Cégeps en tant que générateur de déplacement, on constate une répartition relativement uniforme des montées sur les groupes « pic milieu », « pic 16h » et « pic 17h », ce qui semble plutôt bien refléter la variabilité dans les horaires des étudiants. Enfin, les écoles primaires génèrent pratiquement 70 % des montées (Figure 10) du groupe « pic milieu », tendance pouvant s'expliquer par l'utilisation du transport en commun par les parents après avoir déposé les enfants pour aller au travail, et réciproquement après avoir récupéré les enfants pour regagner le domicile.

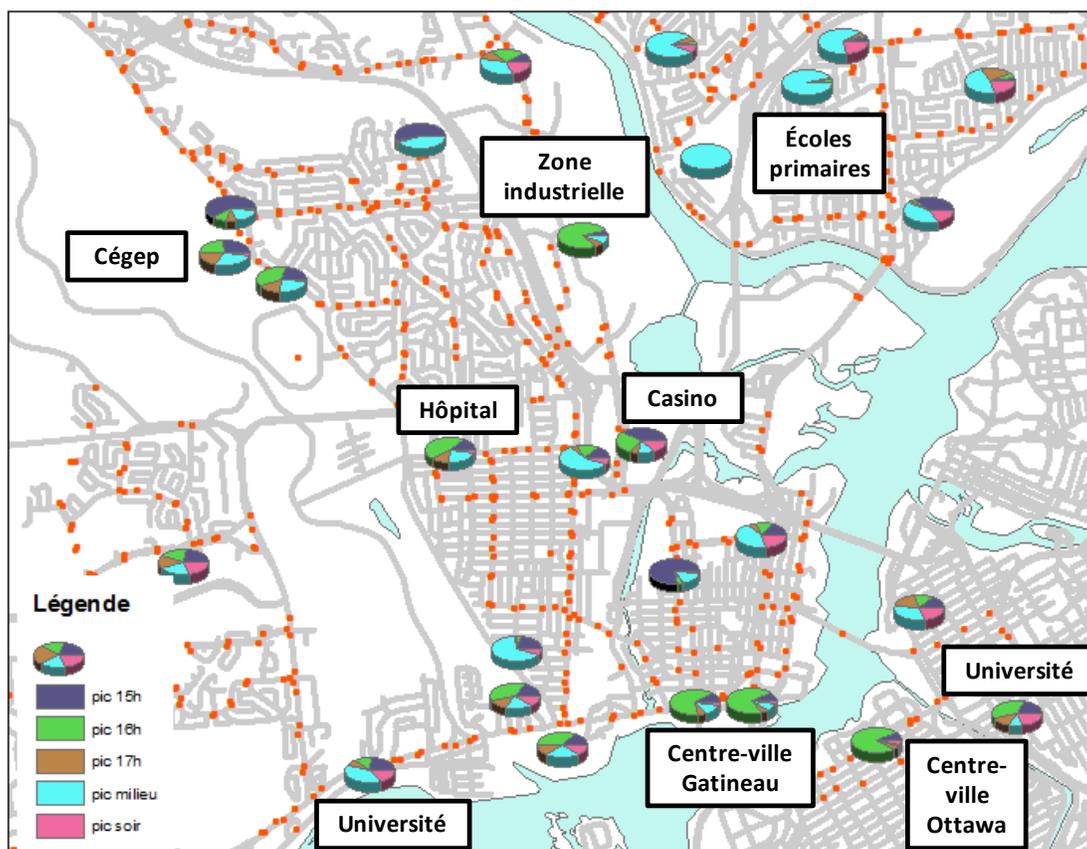


Figure 9: Carte des générateurs selon la répartition des montées par groupe (groupes régulier et du matin omis)

Par ailleurs, on notera que certains générateurs de déplacement induisent plus de 50 % des montées au sein d'un seul groupe sur l'ensemble des groupes influencés. Ainsi, 50 % des montées proches des universités et des édifices gouvernementaux appartiennent au groupe « pic 16h » (Figure 10), 50 % des montées proches des écoles secondaires sont issues du groupe « pic 15h » et plus de 60 % des montées générées par les écoles primaires et les zones commerciales proviennent

du groupe « pic milieu ». Il est intéressant de remarquer que cette dernière tendance se retrouvait également dans la proportion des ALD selon la répartition des montées par groupe (Figure 8). Le comportement ne semble pas uniforme pour tous les générateurs d'une catégorie donnée. Par exemple, à la Figure 10, pour l'ensemble des hôpitaux, le groupe « pic milieu » est prépondérant, tandis que pour l'hôpital illustré à la Figure 9, c'est le groupe « pic 16h » qui domine.

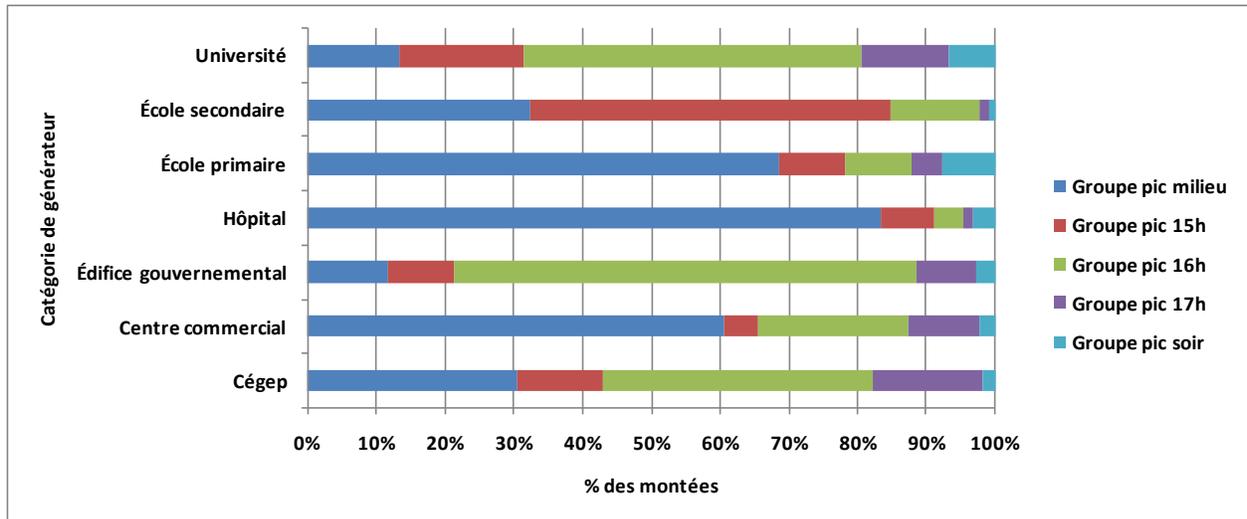


Figure 10 : Pourcentage des montées par catégorie de générateur (groupes régulier et du matin omis)

5 Conclusion

Cet article ouvre ainsi la voie à l'étude de l'impact des générateurs de déplacement sur l'utilisation d'un réseau de transport en commun à partir des données générées par l'utilisation de cartes à puces. La popularisation des systèmes d'informations et de validations des titres (SIVT) couplés aux phénomènes émergents de saturation des réseaux de transport en commun tend à une systématisation de l'analyse des données et ceux dans l'optique de rationaliser leurs fonctionnements. S'intéresser au rôle des générateurs de déplacement permet de voir l'influence d'un pôle géographique sur les heures de pointe et les besoins généraux en transport. Il est cependant important de noter que les groupes ne sont pas stables au cours du temps, en particulier en période de vacances, de fin de semaine et de jours fériés. Néanmoins, il est évident qu'il existe une corrélation significative entre les générateurs de déplacement et la prédominance d'un ou plusieurs groupes. Il serait intéressant d'utiliser la même approche sur un panel important de villes afin de voir s'il existe des tendances marquées par type de générateur. À terme, nous pourrions ainsi

prédire l'influence de l'apparition d'un nouveau générateur de déplacement (par exemple la construction d'un hôpital) sur un réseau de transport en commun, et en adapter ainsi ses capacités.

6 Remerciements

Les auteurs désirent remercier la Société de transport de l'Outaouais pour son soutien aux activités de recherche sur la carte à puce à l'École Polytechnique, ainsi que Clément Piriou, qui a ouvert la voie à l'étude des générateurs via son mémoire de maîtrise.

7 Références

Bagchi, M., White, P.R. (2005). The potential of public transport smart card data, *Transport Policy*, 12, pp. 464-474.

Bonneau, W. And Editors (2002). The role of smart cards in mass transit systems, *Card Technology Today*, June 2002, p.10.

Chapleau, R., Chu, K. K. (2007). Modeling transit travel patterns from location-stamped smart card data using a disaggregate approach, *11th World Conference on Transportation Research*, Berkeley, Californie.

Chu, K. K., Chapleau, R. (2007). Imputation techniques for missing fields and implausible values in public transit smart card data, *11th World Conference on Transportation Research*, Berkeley, California.

Chu, K. K., Chapleau, R. (2008). Enriching Archived Smart Card Transaction Data for Transit Demand Modeling. *Transportation Research Record: Journal of the Transportation Research Board* no. 2063, pp. 63-72.

Clarke, R. (2001). Person location and person tracking: Technologies, risks and policy implications, *Information Technology & People*, 14 (2), 2001, pp. 206-231.

Institute of Transportation Engineers (2003). Trip Generation, 7th edition, ITE, Washington.

Morency, C., Trépanier, M., Agard, B. (2007). Measuring transit use variability with smart-card data, *Transport Policy*, vol. 14, no 3, pages 193-203.

Pirou, C. (2008). MÉMOIRE. *Mémoire de maîtrise en génie industriel (M.Sc.A.)*, École Polytechnique de Montréal, XXX p.

Rakotomalala, R. (2005), "TANAGRA : un logiciel gratuit pour l'enseignement et la recherche", dans les Actes de EGC'2005, RNTI-E-3, vol. 2, pp.697-702.

Trépanier M., Morency C., Blanchette C. (2009), Enhancing household travel surveys using smart card data?, *88th annual meeting of the Transportation Research Board*, Washington, D.C.

Trépanier, M., Chapleau, R., Tranchant, N. (2007). Individual Trip Destination Estimation in Transit Smart Card Automated Fare Collection System, *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, Taylor & Francis, vol. 11, no. 1, pages 1-15.

Trépanier, M., Barj, S., Dufour, C., Poilpré, R. (2004), Examen des potentialités d'analyse des données d'un système de paiement par carte à puce en transport urbain, *Congrès annuel de 2004 de l'Association des transports du Canada à Québec* (Québec), p.4, 10-14.

Trépanier M., Chapleau R., Allard B., Morency C., Trip generator relocation impact analysis methodology based on household surveys, *Institute of Transportation Engineers Journal*, Vol. 73(10), 2003.