Projet Predit Mobilletic : Analyse de la mobilité par les données billettiques

Livrable 3 : Analyse exploratoire des données

Ce livrable est constitué de deux articles publiés :

- Mohamed K. El Mahrsi, Anne-Sarah Briand, Etienne Côme et Latifa Oukhellou, *Utilité des données billettiques pour l'analyse des mobilités urbaines: le cas rennais*. A paraître (2015) dans Données Urbaines, Collections VILLE, Economica.
- Mohamed K. El Mahrsi, Etienne Côme, Johanna Baro, Latifa Oukhellou, *Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data*, In Proceedings of 3rd International Workshop on Urban Computing ACM SIGKDD, New-York 2014.

Utilité des données billettiques pour l'analyse des mobilités urbaines: le cas rennais

Mohamed K. El Mahrsi, Anne-Sarah Briand, Etienne Côme et Latifa Oukhellou
Université Paris-Est, IFSTTAR, COSYS-GRETTIA, F-77447 Marne-la-Vallée, France

De nouvelles sources d'information telles que les données billettiques permettent de proposer d'autres approches pour étudier la mobilité dans les transports en commun. À travers un cas d'étude mené sur Rennes Métropole, nous illustrons comment ces données peuvent être exploitées afin d'extraire des connaissances permettant de caractériser finement les habitudes temporelles des usagers et les profils d'activité des stations composant un système de transport en commun.

Introduction

Le thème de la mobilité intéresse plusieurs communautés de chercheurs en économie, sociologie, urbanisme, géographie, statistiques, etc. Ces dernières années, la croissance démographique, la congestion routière grandissante et les enjeux écologiques liés à la réduction des nuisances environnementales en termes de gaz à effet de serre, de pollution locale et de bruit sont autant de facteurs qui ont motivé l'émergence de nouvelles politiques de mobilité durable. Les pouvoirs publics y jouent un rôle important pour impulser des pratiques de mobilités urbaines visant en particulier à diminuer l'usage des voitures individuelles et à développer l'usage des transports en commun (TC) seuls ou en combinaison avec d'autres modes de mobilité tels que la marche, les vélos partagés, les autos partagés, le co-voiturage, etc.

Une grande majorité des travaux menés jusqu'à présent pour l'étude des mobilités urbaines repose essentiellement sur l'analyse des enquêtes de mobilité. Or nous disposons aujourd'hui d'autres sources de données comme les données billettiques, les traces GSM (téléphone mobile), Wi-Fi ou Bluetooth, la géolocalisation de nos publications sur les réseaux sociaux, etc. générées lors de nos déplacements, lesquelles s'avèrent très riches en termes de traçabilité des mobilités.

La plupart des autorités organisatrices des transports (AOT) ont mis en place des systèmes de billettique avancés qui fournissent une quantité importante de données liées aux déplacements sur l'ensemble des réseaux de transport où ils sont déployés. Ces données possèdent des avantages intrinsèques intéressants comme une certaine exhaustivité, la finesse spatiale et temporelle qu'elles portent, l'absence de biais de réponses que l'on peut rencontrer dans les données d'enquête.
Cet article présente des travaux sur l'utilisation de données billettiques pour l'analyse des mobilités en prenant Rennes comme cas d'étude. Les résultats de telles analyses menées à l'aide d'outils avancés de fouille de données peuvent avoir deux finalités :

- la première est orientée “système de transport”. On souhaite extraire des données billettiques des groupes de stations ayant des motifs d’usage similaires au cours du temps.
- la seconde est orientée “usager”. L’objectif est d’analyser les habitudes des usagers dans leur usage des transports en commun. Le référencement des usagers dans les traces numériques est évidemment anonyme, il n’en permet pas moins de disposer d’une vision longitudinale, intéressante à exploiter en vue de faire émerger des catégories homogènes, de mieux identifier la demande en temps et dans l’espace.

Apport et limites de l’exploitation des données billettiques

Plusieurs travaux ont été menés sur l’utilisation de données billettiques pour l’analyse des mobilités. La mise en place de nouvelles approches de modélisation des mobilités urbaines utilisant des données billettiques soulève néanmoins quelques défis qu’il convient de préciser :

- Les données manquantes : pour des systèmes comme le métro ou le bus, seules les données de l'origine du déplacement sont disponibles (l'usager valide son ticket uniquement à la montée).
- Le volume important de ces données qui fait leur richesse, mais qui pose le problème de leur stockage et de leur traitement.

Ces données permettent d’avoir un flot de données longitudinales. En revanche, aucune donnée socio-économique relative à l’usager n’est disponible. Celle-ci peut être partiellement reconstruite à partir de différentes données estimées comme le lieu de travail, de résidence, le type de titre de transport, etc.

Dès 2004, Bagchi *et al.* s’intérrogent sur le rôle des données billettiques comme nouvelle source d'information pour l'analyse des pratiques de voyage et étudient leur potentiel et leur faculté à compléter, voire remplacer les données plus traditionnelles de type enquêtes. Ils rappellent les avantages de ces données, leurs richesses mais énoncent également leurs limites : pas de validation à l'arrivée, absence du motif des déplacements, échantillon de la population pas nécessairement représentatif.

Une problématique largement traitée dans l’analyse de mobilités au travers de données billettiques est la présence de données manquantes. Tout d'abord la connaissance de la destination finale des déplacements est une donnée importante. Parmi les premiers à s'être intéressés à cette problématique, Barry *et al.* (2002) ont développé une méthodologie qui estime des matrices Origine-Destination (OD) station par station, pour toutes les stations de métro de la ville de New-York. Pour cela, ils déterminent une séquence de validations effectuées tout au long de la journée pour chaque carte de métro. Les auteurs fondent ensuite leur méthodologie sur deux hypothèses, à savoir qu'un grand pourcentage des usagers (i) retournent à la station de destination de leur déplacement précédent pour effectuer le suivant et (ii) finissent leur dernier déplacement de la journée à la station où ils ont commencé leur premier déplacement du jour.

Que l'on dispose de la destination des voyages ou non, la richesse des données billettiques nécessite souvent l'utilisation d'outils spécifiques pour leur exploitation. Agard *et al.* (2006) se sont intéressés à l'utilisation des méthodes de fouilles de données (ou data mining) pour l'étude des données billettiques et à l'utilisation des modèles de planification des transports. Le but est

d'obtenir une représentation améliorée des profils des utilisateurs des transports publics à l'aide d'outils de partitionnement de données comme la classification ascendante hiérarchique ou l'algorithme des K-moyennes (K-means).

Dans leur article de 2008 et dans ceux qui ont suivi, Chu *et al.* ont quant à eux développé une méthodologie visant à enrichir les données billettiques en vue d'obtenir des informations plus complètes à des fins de planification. Cet enrichissement passe par une reconstruction des déplacements des usagers à partir de leurs validations et pas uniquement des origines et destinations. Une attention particulière a été portée à la détection des correspondances (au travers notamment du temps et de la distance entre deux validations pouvant conduire à déduire qu'il s'agit d'un même déplacement). Les auteurs ont également proposé des méthodes visant à reconstituer les chemins spatio-temporels des bus, à détecter et corriger des erreurs dans les données ou encore des méthodes permettant d'associer les usagers à des points d'ancrage (c'est-à-dire des zones géographiques très fréquentées par un même usager et pouvant correspondre à son domicile ou bien à son lieu de travail).

Un autre aspect important des travaux utilisant les données billettiques concerne l'étude des pratiques individuelles de déplacement des usagers. Lathia *et al.* (2010) analysent les données de déplacement individuelles dans le métro londonien (offre de service personnalisé) afin d'estimer les voyages personnels. Ils présentent une méthode de prédiction des heures de voyage personnalisée pour les usagers et font un classement des stations fondé sur les futurs motifs de mobilité. Dans leurs travaux suivants, les auteurs s'intéressent plus particulièrement aux réponses des usagers aux sollicitations de l'opérateur de transport.

L'ensemble de ces travaux montre le potentiel des données billettiques comme nouvelle source d'information pour l'analyse des mobilités. Il met également en exergue un certain nombre de problèmes à résoudre pour que l'exploitation de ces données soit pertinente. Cela peut concerner leur enrichissement, leur traitement ou leur visualisation.

Analyse des données billettiques de Rennes Métropole

Rennes Métropole dispose d'un réseau de transports en commun de taille significative qui couvre 38 communes et dessert plus de 400 000 habitants grâce à plus de 70 lignes de bus régulières et 1 ligne de métro. Dans cette partie, nous présentons la captation et l'enrichissement des données billettiques de ce réseau et nous illustrons l'intérêt des méthodes de fouille de données dans ce contexte pour analyser ces données de grande taille à travers deux études de cas utilisant des algorithmes de clustering. Le clustering est une technique d'apprentissage non supervisé[1] largement utilisée à des fins exploratoires. Il consiste à segmenter un ensemble d'observations en groupes (ou *clusters*) regroupant des observations similaires. L'étude de ces clusters permet de révéler les comportements fréquents et les tendances caractérisant la population étudiée. Nous appliquons deux approches de clustering aux données billettiques issues de Rennes Métropole afin de caractériser son système de transports en commun selon deux angles de vue. Le premier focalise sur les stations de bus et de métro et vise à découvrir des groupes de stations dont les profils d'activité sont similaires (en

[1] L'apprentissage non supervisé regroupe l'ensemble des méthodes d'analyse exploratoire permettant par exemple de trouver des relations et des structures latentes dans des jeux de données, contrairement à l'apprentissage supervisé ces méthodes n'utilisent pas de jeu de données labellisés où les observations sont étiquetées (affectées à des sorties (classes par exemple) bien déterminées).

termes de répartition des validations qui y sont effectuées au cours du temps). Le second point de vue est centré sur les usagers et permet d'extraire des routines temporelles décrivant leurs habitudes vis-à-vis des heures de prise des transports en commun.

Description des données étudiées

Dans le cadre du projet Predit Mobilletic, Keolis exploitant principal de ce réseau de transport urbain a mis à disposition un jeu de données billettiques anonymisé. L'étude exploratoire présentée dans la suite de cet article a été réalisée avec les données du mois d'avril 2014. Celles-ci contiennent des informations sur 5 404 096 validations parmi lesquelles plus de 80 % ont été réalisées par environ 135 000 cartes à puce (appelées cartes KorriGo). Ce jeu de données brutes contient les informations habituelles pouvant être extraites des données billettiques : numéro anonymisé de cartes (pour les validations effectuées avec une carte), date et heure de la validation, station de la validation, type de titre de transport, direction du voyage (pour les bus). Comme dans la plupart des réseaux de transport urbains français, aucune information sur la destination des voyages n'est disponible dans ce jeu de données brutes puisque l'usager n'a pas à valider en sortie du réseau. Pour pallier le manque d'information sur les destinations et détecter les correspondances, les destinations potentielles des validations enregistrées ont été estimées en recherchant la station de la ligne courante la plus proche de la prochaine station de départ de l'usager. Si cette station est située à plus de 500 mètres de la station de prochaine validation, nous n'avons pas affecté de destination. De manière classique les validations en correspondance ont été détectées de la manière suivante : lorsque l'origine et la destination d'une validation sont estimées, le temps de correspondance avec la validation suivante est calculé en retranchant de l'heure de départ de la validation suivante l'heure estimée d'arrivée de la validation courante. Cette heure d'arrivée étant simplement estimée en ajoutant à la date de validation le temps de trajet théorique tel que défini par les tables horaires, disponibles à Rennes au format GTFS (General Transit Feed Specification). Lorsque ce temps est inférieur à 30 minutes, il est raisonnable de considérer que les deux validations forment un unique déplacement parce que liées au même motif. Dans ce cas de figure, les deux validations sont regroupées dans un unique déplacement.

Analyse spatio-temporelle de l'activité du réseau

Pour effectuer une première analyse de l'activité spatio-temporelle du réseau, nous proposons d'utiliser un algorithme de clustering de manière à regrouper les stations ayant des profils d'activité semblables (nombre de validations par heure), même si leurs volumes totaux de validations sont différents, selon une approche déjà présentée par ailleurs (Côme et Oukhellou, 2014). Le jeu de données est tout d'abord nettoyé en supprimant les validations qui ne sont pas affectées à des stations (à causes d'erreurs de géolocalisation) : 5 294 672 validations (98 % du jeu de données original) effectuées dans 686 stations sont retenues. Pour chaque station, les validations correspondantes sont agrégées en comptant le nombre de validations pour chaque paire jour/heure (par exemple, lundi 8h, jeudi 10h, etc.). La description d'une station (notée s_d) pour un jour donné $d \in \{1, \ldots, D\}$ est : $\boldsymbol{s}_d = (s_{d1}, s_{d2}, \ldots, s_{dh}, \ldots, s_{d24})$, avec D le nombre de jours d'observation (30 dans notre cas), $h \in \{1, \ldots, 24\}$ l'heure de la journée et s_{dh} le nombre de validations dans la station pendant l'heure h du jour d.

Un modèle de mélange de lois de Poisson est estimé à partir des comptages de validations dans les stations. Le modèle repose sur l'utilisation de deux ensembles de variables : Z_s qui correspond à un vecteur de variables latentes indiquant l'appartenance des stations à l'un de K clusters (K est fixé a priori) et W_d qui indique le type du jour (jour de semaine ou weekend) pour chacun des jours d'observation. Le modèle suppose que la distribution des stations sur les K clusters suit une loi multinomiale de paramètre π. Connaissant le cluster auquel une station s appartient et le jour d'observation d, les comptages horaires de validations dans celle-ci pendant d sont supposés être conditionnellement indépendants et suivre une loi de Poisson de paramètre $\alpha_s\lambda_{klt}$. α_s est un facteur d'échelle capturant l'activité globale dans la station ce qui permet de regrouper des stations dont les activités ont des silhouettes semblables même si leurs volumes totaux de validations diffèrent et λ_{klt} est un paramètre qui capture la variation du nombre de validations au cours du temps conditionnellement au cluster de la station et au type du jour. Le modèle peut être formalisé à travers l'ensemble des formules suivantes :

$$Z_s \sim M(1, \pi),$$
$$s_{d1} \perp s_{d2} \perp \ldots \perp s_{d24} \mid (Z_{sk}W_{dl} = 1),$$
$$s_{dt} \mid (Z_{sk}W_{dl} = 1) \sim P(\alpha_s\lambda_{klt}).$$

Les paramètres du modèle sont estimés en utilisant un algorithme EM (espérance-maximisation) et le nombre de clusters K adéquat aux données est fixé en appliquant une heuristique de pente avec un critère de maximum de vraisemblance. Pour le jeu de données étudié, ce nombre a été estimé à 14 clusters de stations.

Les profils d'activité au cours du temps des 14 clusters obtenus sont illustrés dans la figure 1 et montrent différents types d'usage des stations. Par exemple, les clusters 1 (fig. 1(a)), 9 (fig. 1(i)) et 14 (fig. 1(n)) sont caractérisés par une forte activité centrée sur la matinée pendant les jours de semaine et une faible activité pendant le weekend. Comme le montre la figure 2 qui présente à la fois la localisations des stations de ces clusters et les densités de population, ces stations se situent au sein des zones densément peuplées de Rennes Métropole (celles appartenant aux clusters 1 et 9 sont localisées exclusivement dans la périphérie de la ville) ce qui suggère qu'elles sont essentiellement utilisées par les usagers se rendant à leurs lieux de travail. D'autres clusters montrent un comportement inverse avec une activité qui se produit essentiellement lors de la seconde moitié de la journée pour les jours de semaine. C'est le cas notamment du cluster 11 où un premier pic d'activité se produit en début d'après midi suivi d'un second pic plus important en fin de journée. Ce cluster est également caractérisé par une très faible activité pendant le weekend. Les stations appartenant à celui-ci sont localisées majoritairement dans les zones d'activité (c'est-à-dire zones réservées à l'implantation d'entreprises telles que les zones artisanales, commerciales, industrielles, etc.) de Rennes Métropole (cf. fig. 3), et sont par conséquent essentiellement utilisées par des usagers effectuant leurs déplacements de retour à domicile à la fin de leurs journées de travail. Un autre cluster intéressant à étudier est celui représenté dans la figure 1(d). Son profil d'usage est très semblable pendant les jours de semaine et le weekend, avec une activité qui croît progressivement jusqu'à la fin de journée pour décroître ensuite. Le positionnement géographique des stations appartenant à ce cluster (non illustré ici) montre que ces dernières sont localisées essentiellement à proximité des centres commerciaux de la ville, ce qui indique

que ces stations sont potentiellement dédiées à des activités de consommation et de loisirs. Le reste des clusters témoignent d'une activité plus « classique » avec notamment deux à trois pics centrés autour des périodes de pointe en jour de semaine et une activité modérée en milieu de journée pendant le weekend.

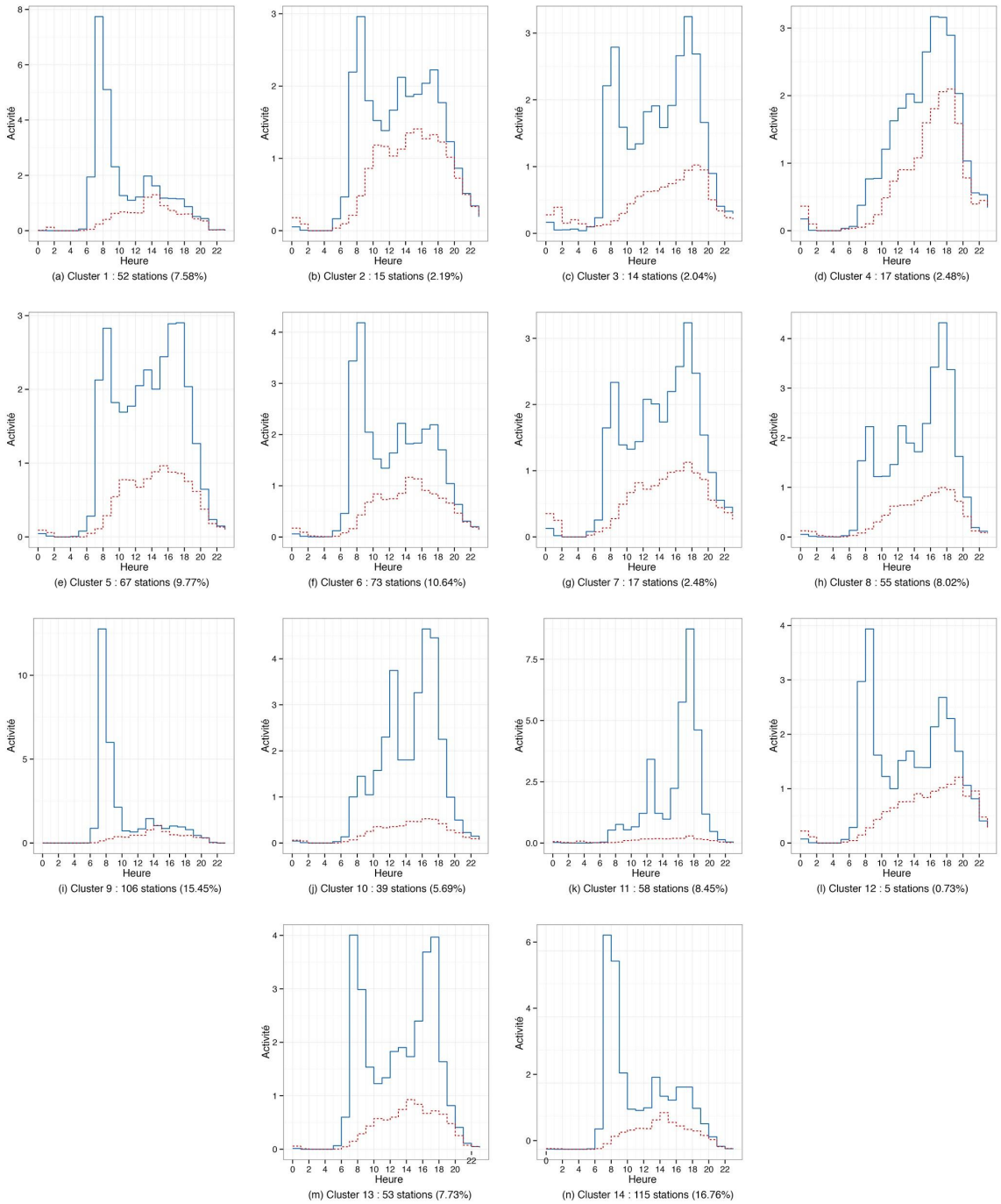


Fig. 1 : Profils d'activité des clusters de stations. L'activité pendant les jours de semaine est indiquée par un trait continu bleu et celle pendant le weekend est indiquée par un trait pointillé rouge. L'échelle de l'axe d'activité est fixée indépendamment pour chaque cluster afin d'améliorer la lisibilité.

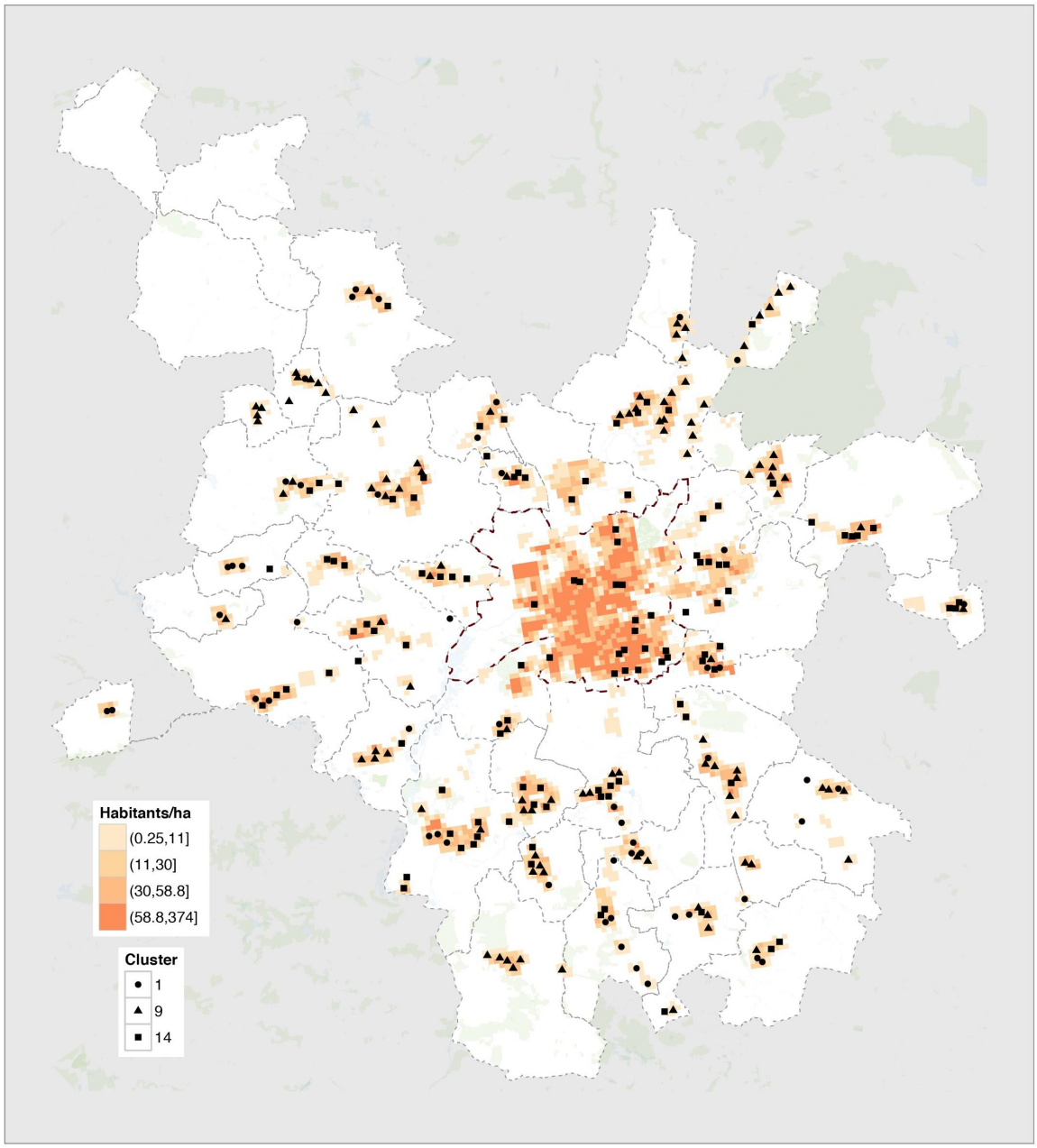


Fig. 2 : Localisation des stations appartenant aux clusters 1, 9 et 14 situés principalement dans les zones résidentielles et caractérisés par une activité se produisant essentiellement en matinée pendant les jours de semaine. Le fond de carte représente Rennes Métropole (la ville elle-même étant indiquée par le contour en tireté) ainsi que les densités de population par carreaux de 200m de côté (Source : INSEE, données carroyées à 200m).

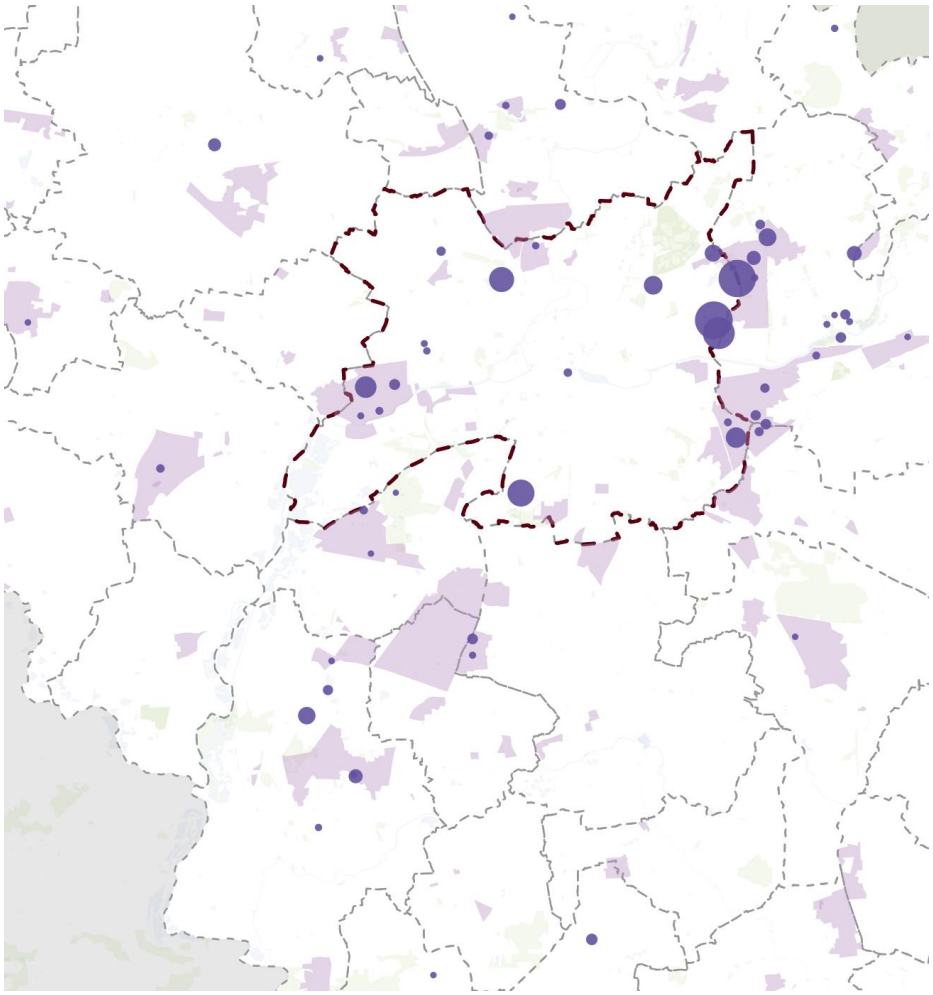


Fig. 3 : Carte des stations du cluster 11. La surcouche en violet indique les zones d'activité à Rennes Métropole. Chaque cercle correspond à une station et les tailles des cercles sont proportionnelles aux facteurs d'échelle (α_s) des stations.

Analyse des routines temporelles des usagers

Nous présentons maintenant un point de vue alternatif sur le système de transports en commun, qui repose sur l'étude des routines temporelles des usagers. L'utilisation du clustering permet en particulier de mettre en lumière des groupes d'usagers ayant des habitudes similaires dans les heures de prise des transports en commun. Les déplacements (après détection des correspondances) de chaque usager sont agrégés sous la forme d'un profil temporel qui décrit le nombre de déplacements qu'il a effectué pour chaque paire de jours de semaine (lundi, mardi, etc.) et d'heure (8h, 9h, etc.). Le profil temporel d'un usager donné peut donc être représenté comme un vecteur $u = (u_1, u_2, \ldots, u_D)$de fréquences (avec $D = 7{\times}24 = 168$ le nombre de paires jour/heure possibles). Ensuite, nous estimons un modèle de mélange d'unigrammes (méthode utilisée dans le traitement automatique du langage pour effectuer le clustering de corpus de documents) à partir des profils temporels des usagers. Dans ce modèle, l'appartenance d'un usager à l'un de K clusters (K étant fixé a priori) est déterminée selon une loi multinomiale de paramètre $\pi = (\pi_1, \pi_2, \ldots, \pi_K)$ qui indique les proportions de clusters. Les N déplacements qu'un usager a effectués sont générés en utilisant une loi multinomiale

conditionnellement au cluster auquel il appartient et qui décrit les probabilités d'effectuer un déplacement pour une heure et un jour donnés. Le modèle est formalisé par les formules suivantes :

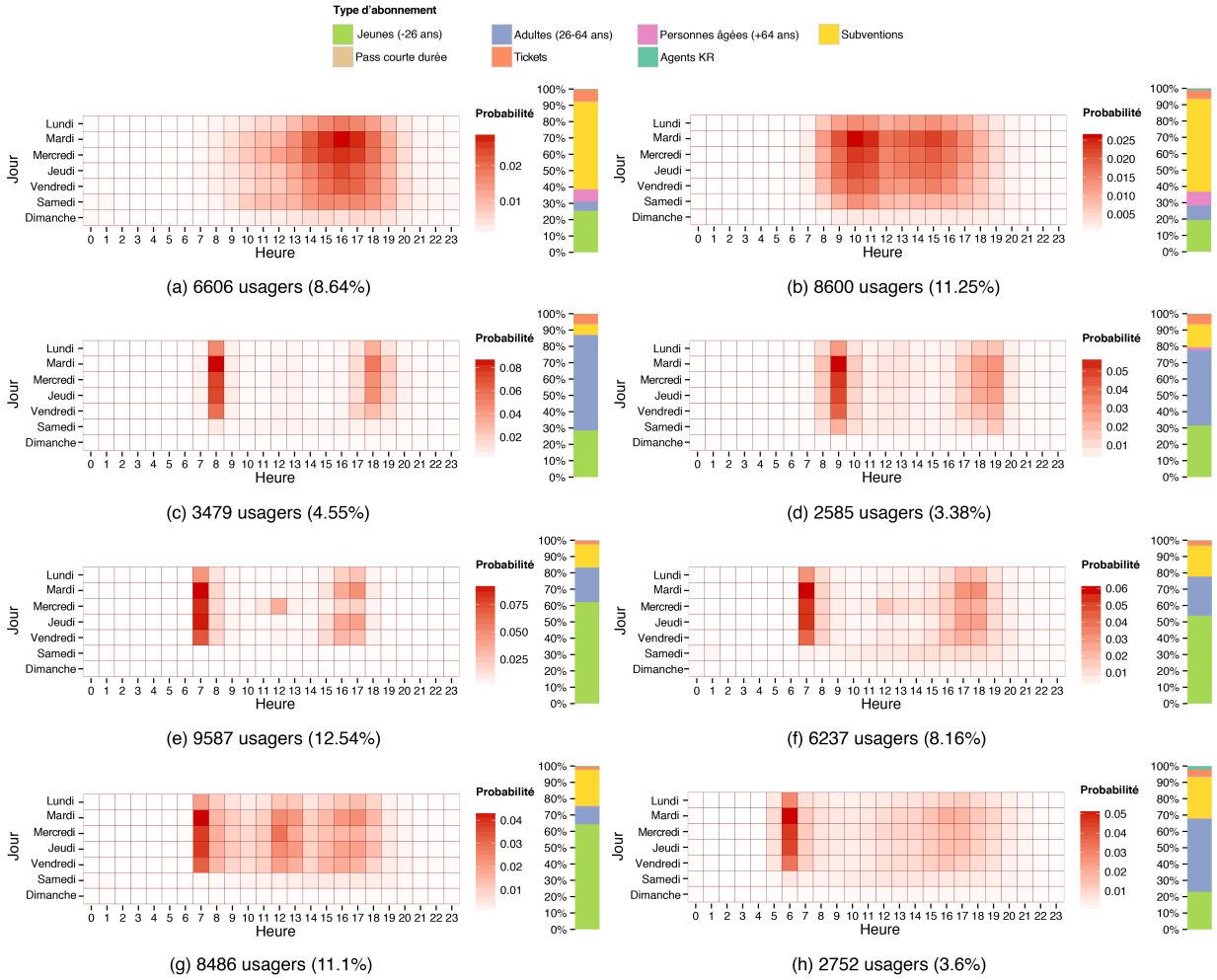
$$z \sim M(1, \pi),$$
$$u \mid z \sim M(N, \beta_z).$$

De façon analogue au clustering de stations, les paramètres π et β du modèle sont estimés grâce à un algorithme VEM (Variationnel EM) et le nombre de clusters est fixé en utilisant une heuristique de pente.

Afin que les clusters découverts soient significatifs, les usagers considérés pour l'étude doivent avoir effectué un nombre suffisant de déplacements. En effet, les usagers occasionnels qui prennent rarement les transports en commun ne sont pas pertinents pour la recherche de motifs et de routines et risquent de bruiter les résultats du clustering. Sur les données billettiques de Rennes, nous retenons seulement les usagers ayant effectué des déplacements pendant au moins 10 jours sur le mois de l'étude, soit 76 478 usagers (56.65% du nombre total d'usagers). L'application de l'approche décrite auparavant sur ces usagers permet de découvrir 13 clusters dont huit sont illustrés dans la figure 4 et montrent différents types de comportement. Par exemple, les clusters que donnent à voir les figures 4(a) et 4(b) sont caractérisés par un usage diffus en après-midi et en matinée respectivement et sont majoritairement composés d'usagers bénéficiant d'une gratuité de transports[2] (accordée selon des critères sociaux, de revenu, etc.). Les clusters représentés sur les figures 4(c) et 4(d) montrent deux pics d'activité en matinée et en fin de journée qui correspondent à des routines typiques de déplacement domicile-travail. Les usagers regroupés dans ces clusters sont majoritairement des adultes possédant un abonnement « normal » et des jeunes de moins de 26 ans. Bien que présentant le même type de comportement, les clusters des figures 4(e) et 4(f) sont également caractérisés par un décalage du deuxième pic de la journée le mercredi. Le fait que ces clusters sont composés essentiellement par des jeunes de moins de 26 ans et qu'en France les cours ne sont pas dispensés mercredi après-midi laissent présager que les membres de ces clusters sont des élèves et des étudiants utilisant les transports en commun pour leurs trajets entre leurs domiciles et leurs lieux d'études. Le cluster de la figure 4(g) montre un troisième pic d'activité en début d'après midi qui vient s'ajouter à ceux de la matinée et de la fin de journée indiquant la présence d'usagers se servant des transports en commun lors de leur pause déjeuner. Enfin, le dernier cluster représenté sur la figure 4(h) est caractérisé essentiellement par un pic qui se produit très tôt le matin comparé aux autres clusters. Notons, que l'approche proposée permet de détecter les subtiles différences entre les routines temporelles des usagers telles que le décalage d'une heure entre les pics d'activité des clusters des figures 4(c) et 4(d) ou encore la présence d'une activité pendant le weekend pour les membres du cluster illustré par la

[2] Les informations sur le type d'abonnement associé à chaque carte KorriGo sont disponibles dans les données sources. À l'origine, l'opérateur de transports maintient une grille de plus de 90 types d'abonnement (basées sur des considérations tarifaires et opérationnelles). Afin de faciliter l'exploitation de cette information, nous les avons agrégé en sept types : (i) Jeunes de moins de 26 ans (43.83% des usagers retenus), (ii) Adultes entre 26 et 64 ans (20.5%), (iii) Personnes âgées de plus de 64 ans (1.89%), (iv) Subventions (personnes profitant de gratuité de transports selon des critères sociaux, etc. représentant 28.84% des usagers retenus), (v) Pass courte durée de moins d'une semaine (0.27%), (vi) Tickets (payement à l'unité par déplacement, 4.12% des usagers retenus) et (vii) Agents KR (regroupant les agents de l'opérateur de transports, 0.54% des usagers retenus).

figure 4(f) comparé à celui de la figure 4(e) où l'usage des transports en commun a lieu exclusivement pendant les jours de semaine.



(a) 6606 usagers (8.64%)

(b) 8600 usagers (11.25%)

(c) 3479 usagers (4.55%)

(d) 2585 usagers (3.38%)

(e) 9587 usagers (12.54%)

(f) 6237 usagers (8.16%)

(g) 8486 usagers (11.1%)

(h) 2752 usagers (3.6%)

Fig. 4 : Quelques clusters d'usagers découverts à partir des données billettiques de Rennes. Les profils des clusters sont illustrés avec des cartes de chaleur indiquant la probabilité d'effectuer un déplacement à une heure et un jour donnés. Les compositions des clusters en termes de types d'abonnements des usagers sont également données.

Conclusions et perspectives

A partir du cas de Rennes Métropole, cet article montre combien l'exploitation des données billettiques peut renforcer la connaissance de la mobilité des personnes. En utilisant des outils avancés d'analyse de données, il devient en effet possible d'identifier des motifs spatio-temporels de déplacement pour mieux comprendre statistiquement l'usage qui est fait d'un réseau de transport, en lien avec les générateurs de déplacement. L'analyse effectuée est construite en croisant deux points de vue : celui du gestionnaire d'infrastructure (focus sur les stations d'un réseau) et celui de l'usager (profils d'usage). Les données billettiques permettent ainsi de disposer d'une vision longitudinale, intéressante à exploiter pour une estimation précise de la demande de manière à anticiper les besoins futurs et planifier au mieux l'offre et le niveau de service. Du point de vue des politiques publiques, l'exploitation des données billettiques peut

servir à mieux anticiper les besoins en matière de transport de manière à impulser des pratiques de mobilités urbaines durables visant à diminuer l'usage des voitures individuelles et à développer l'usage des transports en commun (TC) seuls ou en combinaison avec d'autres modes de mobilité tels que la marche, les vélos partagés, les autos partagés, le co-voiturage, etc.

Ces travaux peuvent être poursuivis en examinant dans quelle mesure les données billettiques peuvent être utilisés conjointement à l'enquête ménage déplacement (EMD), considérée aujourd'hui comme source principale d'information sur la mobilité des personnes. Les deux sources ont une complémentarité évidente par leur construction : une EMD contient un faible nombre d'échantillons et constitue une photographie à un instant donné, avec des méta-informations d'une grande richesse (motifs du déplacement, catégories socio-professionnelles, etc.) ; les données billettiques sont quant à elles nombreuses, d'une grande finesse spatiale et temporelle, mais sont pauvres en méta-informations. De futurs travaux s'attacheront à inventer de nouvelles manières de tirer des enseignements pertinents d'une analyse conjointes de ces sources.

Remerciements

Les travaux présentés dans cet article ont été menés dans le cadre du projet Mobilletic, financé par le PREDIT (Programme de recherche et d'innovation dans les transports terrestres). Nous exprimons notre reconnaissance pour les aides financières reçues dans le cadre de ce projet. Nous tenons également à remercier Keolis Rennes pour avoir généreusement fourni les données qui ont servi à la réalisation de cette étude. Enfin, nous remercions les relecteurs de l'article dont les remarques et les suggestions ont permis d'en améliorer la qualité.

Références bibliographiques

Agard B., Morency C. et Trépanier M., 2006, Mining public transport user behaviour from smart card data, *The 12th IFAC Symposium on Information Control Problems in Manufacturing (INCOM)*, 17-19.

Bagchi M. et White P. R., 2004, What role for smart card data from bus systems, *Proceedings of the Institution of Civil Engineers. Municipal Engineer*, 157, 39-46.

Barry J. J., Newhouser R., Rhabee A. et Sayeda S., 2002, Origin and destination estimation in New York City with automated fare system data, *Transportation Research Record*, 1817, 183-187.

Chu K. et Chapleau R., 2008, Enriching archived smart card transaction data for transit demand modeling, *Transportation Research Record: Journal of the Transportation Research Board*, 2063, 63-72.

Côme E., Oukhellou L., Model-based count series clustering for bike sharing system usage mining: A case study with the vélib system of Paris, *ACM Trans. Intell. Syst. Technol. 5 (3)* (2014) 39:1–39:21. doi:10.1145/2560188.

Lathia N., Froehlich J. et Capra L., 2010, Mining public transport usage for personalised intelligent transport systems, *IEEE International Conference on Data Mining.* Sydney, Australia.

Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data

A Case Study in Rennes, France

Mohamed K. El Mahrsi, Etienne Côme, Johanna Baro, Latifa Oukhellou
Université Paris-Est, IFSTTAR, COSYS-GRETTIA
F-77447 Marne-la-Vallée, France
mohamed-khalil.el-mahrsi@ifsttar.fr, etienne.come@ifsttar.fr,
johanna.baro@ifsttar.fr, latifa.oukhellou@ifsttar.fr

ABSTRACT

Data collected by Automated Fare Collection (AFC) systems are a valuable resource for studying the travel habits of large city inhabitants. In this paper, we present an approach to mining the temporal behavior of the passengers in a public transportation system in order to extract relevant and easily interpretable clusters. Such classification can be useful for several applications. It may help transport operators better know the demand of their customers and propose targeted incentives, services, and tools accordingly. From a city perspective, this may also help redesign and improve existing transportation policies. To achieve this objective, an additional step of analysis is required and the clustering results need to be contextualized. The spatial location of the different types of passengers is then of great interest. We propose a first step in this direction through a rough estimation of the regular passengers' "residence" location and the analysis of socioeconomic information available at a fine-grained spatial level. The approach is applied on a real dataset from the metropolitan area of Rennes (France) with four weeks of smart card data containing trips made by both bus and subway.

Categories and Subject Descriptors

I.5 [**Pattern Recognition**]: Clustering; H.2.8 [**Database Management**]: Database Applications—*Data mining*

Keywords

smart card, public transportation, clustering, topic models.

1. INTRODUCTION

Presently, Automated Fare Collection (AFC) systems are widely adopted by public transportation operators in large cities and metropolitan areas. These systems are based on contactless smart cards that passengers credit and use when boarding buses or accessing subway stations. While the main purpose of AFC systems is to manage revenu collection, they also offer the opportunity to study the day-to-day travel habits of passengers using the transit network. In fact, each transaction registered when a smart card is used contains fine-grained information not only about the passenger's identity but also about the time, boarding location, and —in some cases— the boarded bus or subway line. This makes it possible to reconstruct the detailed profiles of the passengers over long periods of time and analyze them to study various aspects such as travel behavior and regularity, multi-modality, turnover rates, etc.

Understanding the passengers' travel behavior can be useful in a variety of applications. For instance, it can be used to assess the performance of the transit network, detect irregularities (e.g. frauds, defective equipment, etc.), forecast demand with more accuracy, make service adjustments that accommodate with variations in riderships across weekdays and seasons, etc. From a city perspective, smart card data can also prove valuable to evaluate the public transportation accessibility level, detect if particular communities or areas are underserved, and react accordingly to consolidate and improve the existing transportation policies. In this context, however, smart card data do present some shortcomings: due to privacy concerns, the collected data are anonymized and personal information about the individuals (e.g. gender, age, address, income, etc.) are not available. Consequently, inferring the socioeconomic characteristics of the passengers can prove to be difficult.

In this paper, we demonstrate how smart card data can be used conjointly with socioeconomic data to understand passenger behavior in public transportation. The contributions of this work are the following:

- We study the extraction of travel patterns from smart card data. To this effect, we construct temporal passenger profiles based on boarding information and apply a generative model-based clustering approach to discover groups (or clusters) of passengers who behave similarly w.r.t. their boarding times.
- We study how passenger travel habits relate to socioeconomic characteristics. To do so, we start by assigning passengers based on their boarding information to "residential" areas that we established through a clustering of socioeconomic data of the city. Then, we inspect how socioeconomic characteristics are distributed over the passenger temporal clusters.

UrbComp'14, August 24, 2014, New York, NY, USA.

- We apply our approaches on socioeconomic data and a real smart card dataset covering four weeks of boarding validations registered in the metropolitan area of the city of Rennes (France).

The rest of this paper is organized as follows. Related work is discussed in Section 2. We present our approach to clustering smart card data in order to study passenger temporal habits in Section 3. Our attempt to pair smart card and socioeconomic data is reported in Section 4. Finally, we conclude the paper in Section 5 with general remarks and future research directions.

2. RELATED WORK

Bagchi et al. [2, 3] were among the first researchers to substantiate the potential of smart card data for transit planning. The authors discuss the advantages and shortcomings of smart card data w.r.t. their capability of capturing information about passenger trips and present case studies to illustrate how rules-based processing can be used to infer turnover rates, trip rates, and detect linked trips from the collected data. The authors also emphasize that since some information are not captured by smart card data (e.g. journey purpose, satisfaction w.r.t. the transport service, etc.), the latter cannot entirely supersede existing data collection approaches (mainly direct surveys) but rather complement them.

Morency et al. [16] conduct an analysis of the variability of travel behaviors in public transit based on bus trip data. Trips are aggregated into transactions, each representing the daily profile of a given smart card on a given date. A transaction is represented using 27 attributes: a smart card id, a date, a type of day (e.g. D1-Monday, D2-Tuesday, etc.), and a feature per hour indicating if the smart card was used for validation or not (e.g. feature H10=1 indicates that the smart card was validated at least once between 10:00 and 10:59). First, the authors conduct analyses in order to calculate global indicators of regularity based on the activity rate (i.e. ratio between the number of days where a smart card was validated for boarding and the number of days of observation), the number of boardings per day, and the number of different boarding stations. Later on, the authors focus on studying the regularity of the individual behavior of two cards (an elderly and a regular adult). Two aspects are studied: (i) the number of boardings per day and (ii) boarding time. For the latter, k-means clustering is applied in order to identify (separately for each user) clusters of similar days w.r.t. boarding times. A similar analysis of weekly travel behavior using the same data from [16] is conducted by Agard et al. [1]. Here, however, the bus trips are aggregated into weekly transactions each describing the 5 weekday activity of a passenger during a given week. A transaction comprises a smart card id, the fare type of the smart card, the week, and 20 binary features representing 5 weekdays with 4 periods (AM = 5:30 to 8:59, MI = 9:00 to 15:29, PM = 15:30 to 17:59, and SO = 18:00 and over) per day. Hierarchical Agglomerative Clustering (HAC) and k-means are applied to the transactions in order to study group behavior. For instance, the authors showcase how inspecting the composition of clusters w.r.t. fare type and the evolution of the latter over the studied weeks helps discover groups such as "typical workers" (typically traveling during AM peaks and PM peaks during weekdays) and atypical behaviors (e.g. a shift of regularity in students' behavior due to spring-break).

Lathia et al. [12] discuss how smart card data can be used to reveal hidden individuals' behaviors and study their response to travel incentives. First, a comparison between an online survey's results and real travel data collected from Transport for London's (TfL) Oyster smart card system is presented. This comparison attempts to characterize the difference between the perceived and the actual behavior of passengers by studying various aspects such as trips per day frequency and regularity, atypicality of travel modality and origin and destination stations, as well as cash-fare purchasing habits. Secondly, the authors demonstrate how smart card data can help study the extent to which incentives (such as time-varying fares, price capping, etc.) proposed by transport operators influence passengers' decisions.

In [21], the behavioral differences between travel card holders and Pay as you go passengers in the London Underground are studied using smart card data over a four months period. Fuse et al. [11] investigate the use of bus smart card data to determine useful information (such as travel time and number of passengers) that can be used for congestion spot analysis and the improvement of bus stops planning. In [15], a data mining approach to extracting individual transit riders' travel patterns and assessing their regularity from a large smart card dataset with incomplete information is presented. The study focuses on two aspects: (i) characterizing the spatial and temporal travel patterns of transit riders on an individual basis, and (ii) determining the regularity of these patterns. The authors use smart card data collected from Beijing's (China) bus and subway AFC systems (offering flat-rate fares and distance-based fares). The adopted methodology is composed of three steps: (i) first, each passengers transactions are retrieved; (ii) the spatiotemporal relationships between these transactions are used to generate trip chains on which (iii) data mining techniques are applied in order to extract the passenger's travel patterns and regularity information.

Inferring "community well-being" from smart card data is discussed in [13]. First, stations are mapped, based on geographic proximity, to communities and IMD (Index of Multiple Deprivation) scores obtained from national census results. Then, trip data are used to compute a station-by-station flow matrix representing locations visited by different communities. The station-to-IMD mapping and station-to-station flow matrix are used to study the correlation between IMD and flow and calculate homophily indices. Results suggest that, contrary to well-off areas attracting people from communities of varying social deprivation, deprived areas do suffer from social segregation and tend to attract people only from other deprived areas. To some extent, this work is similar to what we present in Section 4 as it involves using socioeconomic data. In [13], these are used in order to study how communities relate to each other, whereas in our context we try to study if and how they influence temporal behavior of public transportation passengers.

Personalizing transport information services based on smart card data is discussed in [14]. Among their contributions, the authors use clustering to prove that usage of public transportation can vary considerably between individuals. Each passenger's trips are aggregated into a weekday profile describing his temporal habits (four time bins are used: early morning, morning rush, day time, and evening rush) and hierarchical agglomerative clustering is used to discover groups of passengers with different habits (e.g. commuting

regularly during both rush hours, exclusively in the evening, etc.). Contrary to the approach we present in Section 3, all weekdays are treated equally which might cause day-specific behaviors to be ignored.

Other relevant references in the field include [18, 8]. The interested reader is also referred to the survey in [22] in which the use of public transit data to improving transportation systems is discussed, as well as the thorough literature review in [19].

One key difference between the approach reported in this paper and previous work involving clustering smart card data is the nature of the involved clustering algorithm: most existing approaches to clustering smart card data (e.g. [16, 1, 8]) rely on the use of similarity-based clustering algorithms that minimize Euclidean distortions. These approaches are often criticized for yielding poor results [20]. Our approach, in contrast, relies on the use of a generative model-based clustering (mixture of unigrams). Additionally, some existing approaches consider either clustering the trips of each passenger individually [16, 15] or disregard the identities of the passengers [8, 1]. In our approach, similarly to [14], we consider each passenger (and trips pertaining to him) as a single observation and conduct the clustering accordingly.

3. CLUSTERING PASSENGERS BASED ON TEMPORAL PROFILES

In this Section, we present the smart card dataset that we use for our study and give the details of our approach to clustering public transportation passengers based on their temporal habits.

3.1 Case Study Dataset

We conduct our study on smart card data collected through the automated fare collection system of the Service des Transports en commun de l'Agglomération Rennaise (STAR). STAR operates a fleet of 65 bus lines and 1 subway line serving the metropolitan area of Rennes, which is the 11^{th} largest city in France with more than 200000 inhabitants. The operator established its automated fare collection system on March 1^{st}, 2006 and offers the possibility to travel on the network using a KorriGo smart card. As of 2011, 83% of the trips in STAR's transportation system were made using a KorriGo card.

The original dataset spans over the period from March 31^{st}, 2014 up to April 25^{th}, 2014, and contains both bus and subway boarding validations from more than 140000 KorriGo card holders. Each card is assigned a unique, anonymized identifier in order to ensure privacy and no personal information about the passengers were made available to us. Each validation contains, in addition to the card identifier, the date and time of the operation, the name and identifier of the boarding station, the name and identifier of the boarded bus or subway line, as well as information about the card type.

For our study, we only retain 53195 "regular" passengers. In order for a given smart card holder to qualify as a regular one, we check two conditions: the passenger (i) must have used his smart card for at least ten days during the studied period, and (ii) must have made his first boarding after 4 am each day at the same station 50% of the time (in which case the station is assigned as his "residential" station). The purpose of this step is to enhance the clustering results by removing occasional passengers for which the number of observed trips is insufficient and to have a rough estimate of the spatial location of the user residences. This is of interest since we also have access to a fine description of the socioeconomic characteristics of the city inhabitants through data released on a regular grid containing 200m per 200m cells by the French National Institute of Statistics and Economic Studies (INSEE).

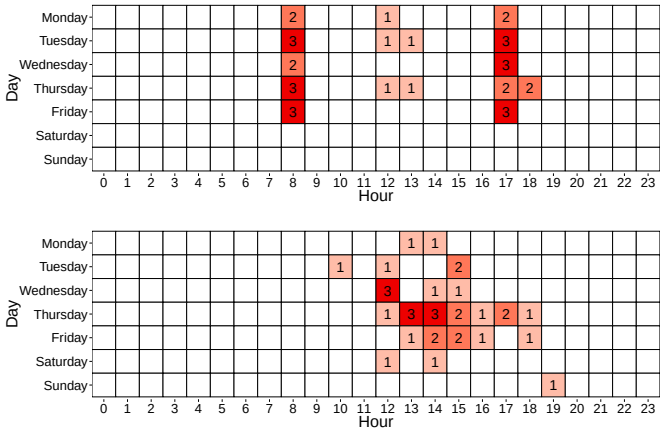


Figure 1: Temporal profiles of two passengers sampled from the smart card dataset.

3.2 Methodology

The aim of this part of our study is to discover groups of passengers who exhibit similar behaviors from a purely temporal standpoint (i.e. passengers taking public transportation at the same times without accounting for the boarding locations). Intuitively, the discovery of these groups can help identify frequent patterns in the way passengers use public transit and characterize the demand accordingly.

We start by aggregating each passenger's validations into a "weekly profile" describing the distribution of all his trips over each hour (0 through 23) of each day of the week (Monday through Sunday). Therefore, each passenger is an observation over 168 variables: the first variable is the number of trips he took on Monday 0 to 1 am, the second is the number of his trips on Monday 1 to 2 am, and so on. The temporal profiles of two passengers are illustrated in Figure 1 and we denote one such profile by $\mathbf{u}$.

Next, we cluster the passenger profiles. We do so by estimating a mixture of unigrams model [17] from our data, an approach often used to cluster documents in the context of information retrieval. Under this perspective, a passenger can be regarded as a "document" containing a collection of "words", with each word being, in our case, a combination of a day and an hour (e.g. Friday 10 am). In a (simple) unigram model, the words of each document are drawn independently from a single multinomial distribution. In a mixture of unigrams model, each document is first generated by selecting a cluster (called topic), then the words of that document are drawn from the conditional multinomial distribution relative to that specific cluster. Each cluster is thus described concisely as a distribution over the vocabulary. In our context, this translates to a cluster being the description of when trips

are more or less probable to be made. More formally, we use the following generative model for the data:

$$z \sim \mathcal{M}(1, \pi) \ , \quad (1)$$

$$\mathbf{u}|z \sim \mathcal{M}(D, \beta_z) \ . \quad (2)$$

with π the clusters' proportions, D the number of displacements recorded in $\mathbf{u}$ and β the cluster profiles. The parameters π and β were estimated from the data using a classical Expectation-Maximization (EM) algorithm. This maximizes the likelihood of the profiles which is derived from their distribution given by:

$$p(\mathbf{u}) = \sum_z p(z) \prod_{d=1}^{D} p(u_d|z) \ .$$

The approach was compared with possible alternatives that also lead to a compact representation of the passenger temporal profiles such as the Latent Dirichlet Allocation (LDA) [7], but eventually we chose to keep this simple approach for the ease of interpretation of the clustering results.

Another concern when estimating a mixture of unigrams model is that the number of clusters K must be specified in advance. In order to retrieve the model that best fits our data, we first use an EM algorithm to estimate the mixture of unigrams models while varying K from 2 to 30. To select an appropriate number of clusters, penalized likelihood criteria such as the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are widely used and asymptotically consistent but they are also known to be less efficient in practical situations than on simulated cases. To overcome this drawback in real situations, [6] have recently proposed a data-driven technique, called the "slope heuristic", to calibrate the penalty involving penalized criteria. The slope heuristic was first proposed in the context of Gaussian homoscedastic least squares regression and was since used in different situations, including model-based clustering. Birge et al. [6] proved the existence of a minimal penalty and that considering a penalty equal to twice this minimal penalty allows to approximate the oracle model in terms of risk. The minimal penalty is estimated in practice by the slope of the linear part of the objective function with regard to the model's complexity. A detailed overview and advices for implementation are given in [4].

3.3 Results

By proceeding as described earlier, we obtain a set of 16 clusters —that we refer to as "temporal clusters" from now on— each describing a temporal mobility pattern. In order to analyze the results, we can inspect visual representations of the daily temporal profiles obtained for each cluster (cf. Figure 2) and cross these visualizations with the proportions of fare types used by the passengers belonging to the corresponding clusters (cf. Figure 3).

As it can be seen on Figure 2, a first distinction can be made between two types of clusters: (i) clusters exhibiting routine behaviors (clusters 5 up to 16, except cluster 6), and (ii) clusters where the transport demand seems to be diffuse and does not portray typical home-to-work commutes during weekdays (clusters 1 up to 4). For example, one can clearly see that clusters 11, 12, and 14 are characterized by a morning peak of use at 7 am (resp. 8 am for cluster 14), an evening peak at 4 pm (resp. 5 pm for cluster 11), and a rush of usage at Wednesday 12 am (Wednesday is the day-off for

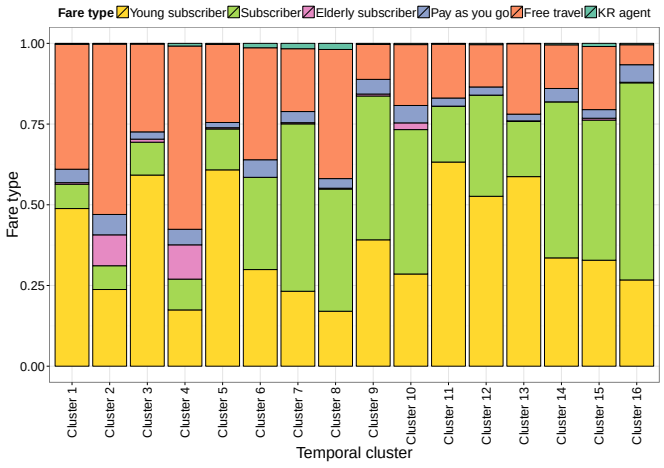


Figure 3: Proportions of fare types per temporal cluster.

students). Cluster 13 shows a quite similar behavior since it also highlights a morning peak at 7 am, but the usage on Wednesday and the evenings is more diffuse. We can also notice that for clusters 11, 12, and 14, no trips were made during the week-end, whereas cluster 13 exhibits a Saturday usage. Regarding the proportions of fare types per temporal profile of each cluster depicted in Figure 3, we can see that most of the passengers in clusters 11, 12, and 13 are made by " young subscribers", the difference in traveling behavior on Wednesday and evenings can be explained by the difference on daily commuting habits between school children and college or high school students for which leaving hours are more flexible. However, the proportion of "young subscribers" is less important in cluster 14 compared to the other clusters. This suggests that most of these passenger habits are indeed linked to daily school schedules but are rather made by parents aligning their trips with the school-leaving hours of their children.

Considering the remaining clusters (9, 10, 15 and 16) portraying daily routine behaviors, varying kinds of typical commuting patterns can be identified. Cluster 15 is related to passengers who travel in the morning peak (at 8 am), in the lunch break between 12 pm and 2 pm, and again in the evening rush hour between 5 pm and 6 pm. A quite similar behavior can be observed for passengers in cluster 16 which, in contrast, do not use transit during the lunch break. Regarding the proportions of fare types, " young subscribers" are largely represented in this cluster. Clusters 9 and 10 also exhibit a morning peak of usage at 9 am and evening peak at 6 pm and 7 pm (one-hour shift compared to Cluster 16). One of the main differences between these two clusters is the fact that passengers in cluster 10 use transit during Saturday more frequently than those in cluster 9.

Clusters 7 and 8 exhibit an early morning use (at 6 am for cluster 7, at 5 am for cluster 8) and no transit use during the week-end, which suggests that the trips are not made for leisure but are rather exclusively related to employment. The free travel fare type is higher for cluster 8. Cluster 5 exhibits daily routine travels. The peaks of transit use are early in the morning at 7 am, in the lunch break and in a

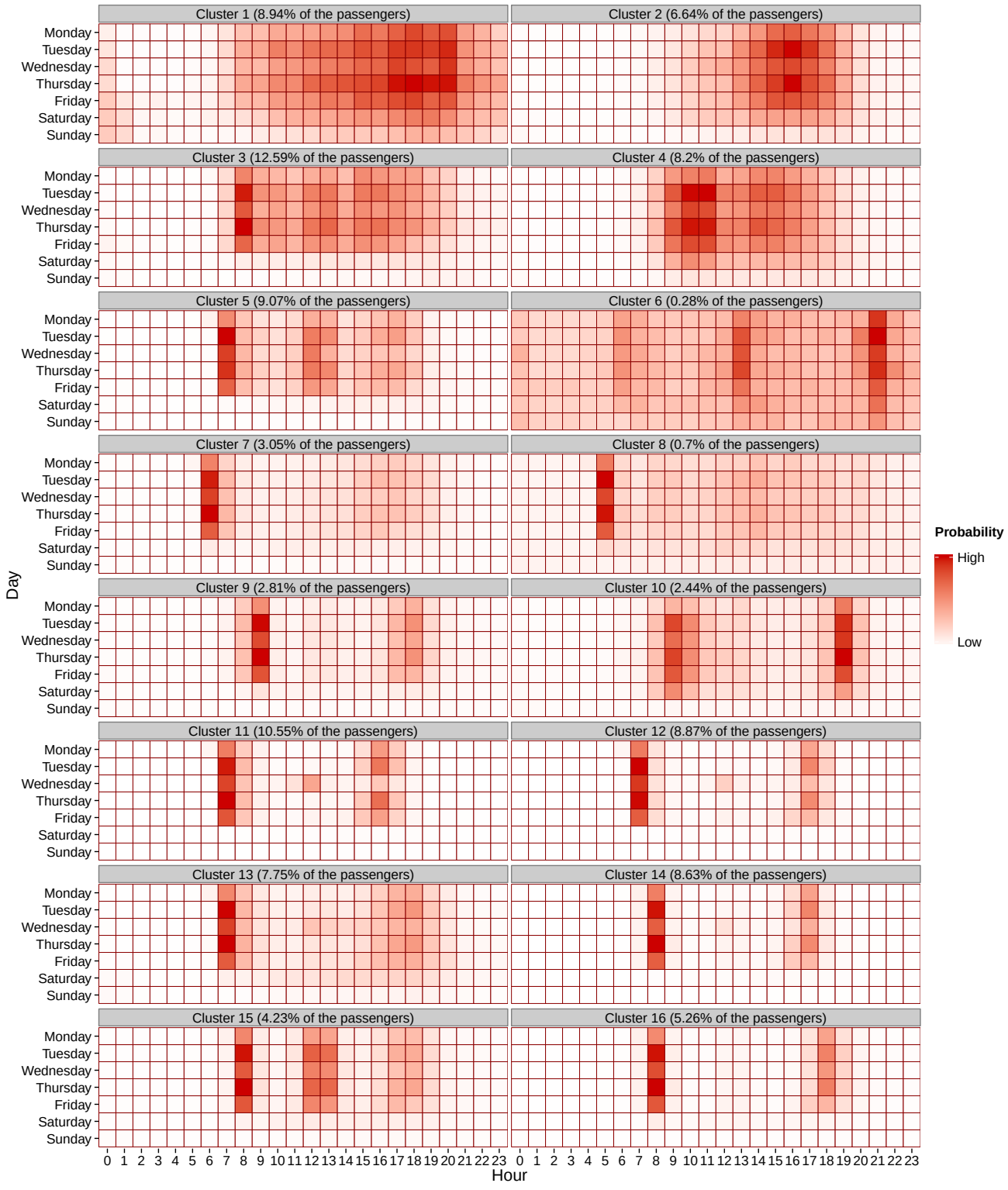


Figure 2: Temporal clusters estimated using the mixture of unigrams model.

diffuse manner in the evening. Passengers in this cluster do not use the transit in the week-end neither. Crossing these remarks with the fact that most of these users are "young subscribers", one can conclude that most of the passengers of this cluster are students.

Clusters 2 and 4 do not exhibit daily routines, meaning thus that such passengers travel after the morning rush hour and in a sporadic manner during the evenings. As it can be seen on Figure 3, these two clusters contain the highest proportions of elderly subscribers and passengers benefiting from free travel.

Finally, clusters 1 and 6 exhibit a diffuse daily usage of the transit system. However, they differ since some peaks of usage can be identified on the temporal cluster 6 such as 6 am, 1 pm, and 9 pm. Also, the proportions of fare types in this cluster are balanced between young subscribers, free travel and subscribers whereas cluster 1 mainly includes "free travel" and "young subscribers". The distinctive travel pattern in this cluster suggests that the passengers are laborers that work in rotating shifts.

4. ADDING CONTEXT THROUGH SOCIO-ECONOMIC DATA

We now discuss how socioeconomic data can be used to complement the study of temporal clusters.

4.1 Clustering of Socioeconomic Variables

In order to improve the analysis of the temporal clusters of passengers, we propose to introduce contextual information regarding the socioeconomic characteristics of the places of residence assigned to passengers through their "residential" stations. Socioeconomic data released on regular grids containing 200m per 200m cells by the French National Institute of Statistics and Economic Studies (INSEE) are available to conduct this study. The data contain several variables such as population count, income per consumption unit, cumulative living area of the households, etc.

Our intuition is to perform spatial clustering over these socioeconomic variables in order to obtain a synthetic description of the neighborhoods of the stations assigned to passengers. We address this spatial clustering problem using a Hidden Random Markov Field (HRMF) model that is well adapted to processing spatial data. The particularity of spatial data is that they cannot be viewed as a collection of independent variables because the values of the neighboring spatial entities are very often correlated one with another. HRMF models are often used for image segmentation or pixel-based classification [5] because they allow the incorporation of spatial dependencies in the building of a classifier through a neighboring system encoded by a Potts model for example. Spatial clustering in this context involves the estimation of the parameters of a hidden Markov field that corresponds to a discrete variable encoding the K clusters, and that must be discovered from the observation of the socioeconomic variables (supposedly following a multivariate Gaussian distribution). To estimate these parameters, several solutions exist based on the use of an EM-type algorithm. Notably, [9] proposed the NREM algorithm based on the mean-field approximation. It is designed for the parameters estimation of a HRMF in an unsupervised learning framework. Being able to fully estimate the parameters, including β the symmetric matrix of the Potts model encoding the spatial interactions between clusters, is an advantage that encourages us to use this particular algorithm because it gives more flexibility to recover the underlying clusters and since the clusters that we aim to identify may not have the same frequencies of apparition, they can have more or less grouped spatial structures. In the rest of this paper, the clusters retrieved as aforementioned will be referred to as "socioeconomic clusters".

Table 1: Mean and standard deviation of the socioeconomic variables in the socioeconomic clusters ($K = 7$).

Cluster	Area (m^2)	Income (€)	Population
Collective housing, Medium+ income	15696 (7763)	21889 (1791)	392 (194)
Collective housing, Low income	12510 (7553)	14499 (2364)	397 (244)
Individual housing, High density	8495 (2452)	20804 (2502)	239 (71)
Individual housing, High income	4359 (1847)	24422 (1899)	99 (45)
Individual housing, Medium inc. & dens.	3721 (1576)	20761 (2379)	103 (45)
Individual housing, Low density	695 (337)	21655 (2794)	17 (9)
Individual housing, Low density	167 (85)	21361 (2749)	4 (2)

N.B. standard deviation is indicated in parentheses.

We run the NREM algorithm on a set formed by the cells of the grid containing one or more stations and their neighboring cells using 8-D connectivity. This way we have a spatial coverage large enough to characterize the living neighborhood of the passengers, assuming here that passengers travel only a small distance from their home to the stations (less than 600 meters). We run the algorithm with a 7-colors Potts model. We choose here the number $K = 7$ manually according to the best interpretability of the clusters. Model selection based on log-likelihood criteria are difficult to apply here because we can only compute an approximated version of the log-likelihood and the existing approximated criteria [10] show inconsistent behaviors in our application. The mapping of the clustering results are visible on Figure 4. On this figure and for the crossing of the results with the temporal clustering, two clusters have been aggregate together because of their very similar profiles (cf. Table 1). The six remaining clusters provide a characterization of the residential neighborhoods in terms of level of income, population density, and type of buildings deduced from the combination of population and build area densities. We can observe from the mapping of the result that the positions of these clusters seem related to the distance to the center of the city of Rennes, which could have an influence on the analysis of the clusters describing the temporal behavior of the transportation operator's passengers.

4.2 Results

We start by examining the proportions of fare types used by the passengers of each of the socioeconomic clusters retrieved through clustering of socioeconomic variables (cf. Figure 5). The figure reveals some predictable results. For instance, we note that the highest proportion of passengers benefiting from free travel (which is attributed based on social and income considerations) is observed for the socioeconomic clus-

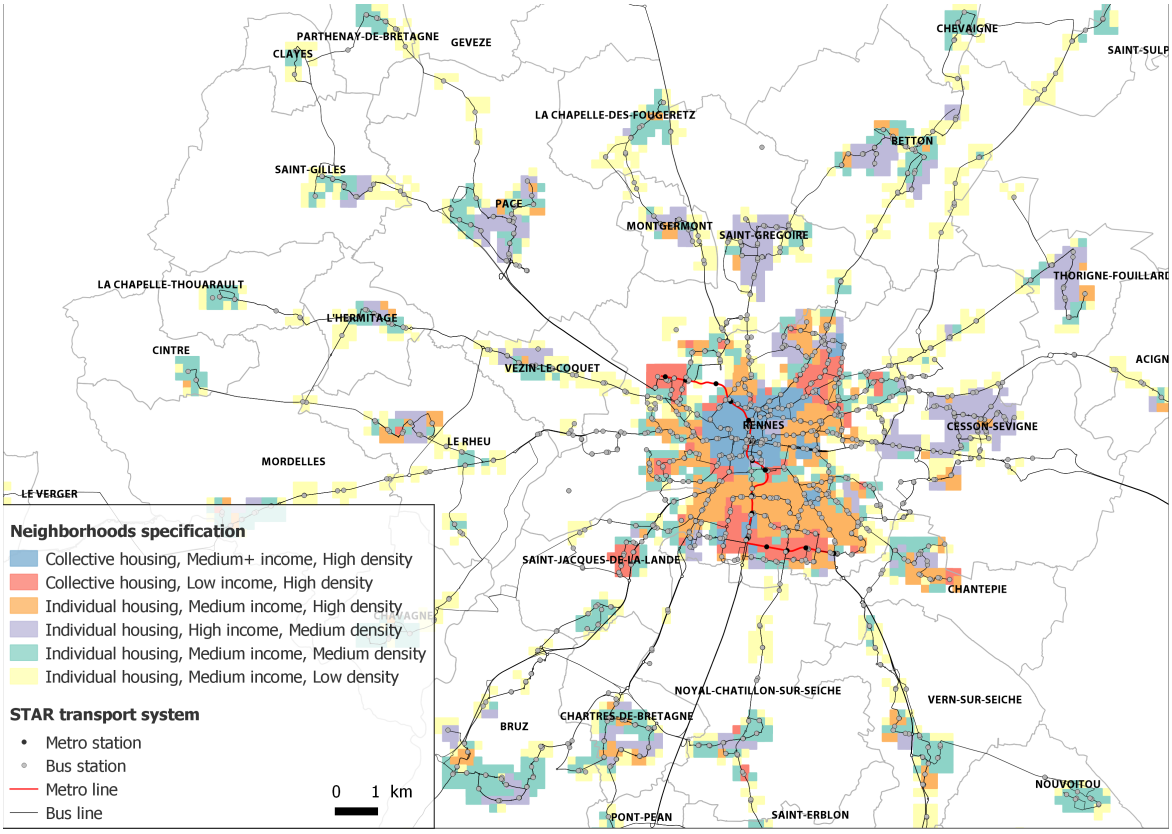


Figure 4: Socioeconomic clustering of the residential neighborhoods.

ter "Collective housing, low income, high density", whereas the lowest proportion of this fare type is in the "individual housing, high income medium density" cluster. The lowest proportion of young subscribers is also observed in the "Collective housing, low income, high density" cluster, which is also expected since —inter alia— students in this group logically use free travel instead of student-targeted fare types. The distribution of elderly subscribers seems balanced between clusters, except for the cluster of "Individual housing, Medium income, Low density" from which they are practically absent. This suggests that elderly people who used the public transport network live in medium to high density areas.

We also attempt to identify links between the socioeconomic clustering of the city and the temporal clustering performed on passengers profiles. To this effect, we inspect the proportions of socioeconomic clusters for each temporal cluster (cf. Figure 6). This figure shows that users belonging to cluster 8 which is characterized by an atypical early morning peak (5 am) are mainly located in "collective housing, low income, high density" areas, thus suggesting that this transport demand is made by early workers located in this kind collective housing (the corresponding spatial location can be seen on Figure 4). The same observation can be made for the temporal cluster 6 characterized by three peaks of usage at 6 am, 1 pm, and 9 pm. These peaks can correspond to laborers working nightshifts or daytime schedules involving three working teams.

Overall, the socioeconomic clusters seem to be similarly distributed across all temporal clusters. The highest demand

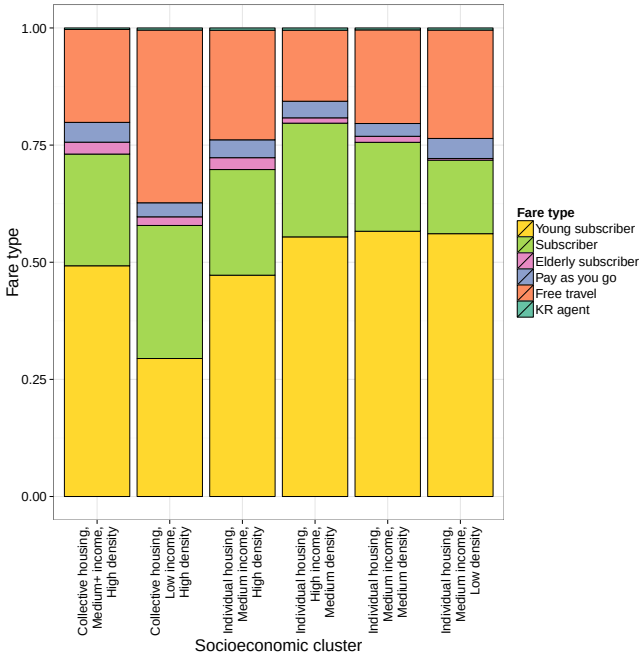


Figure 5: Proportions of fare types per socioeconomic cluster.

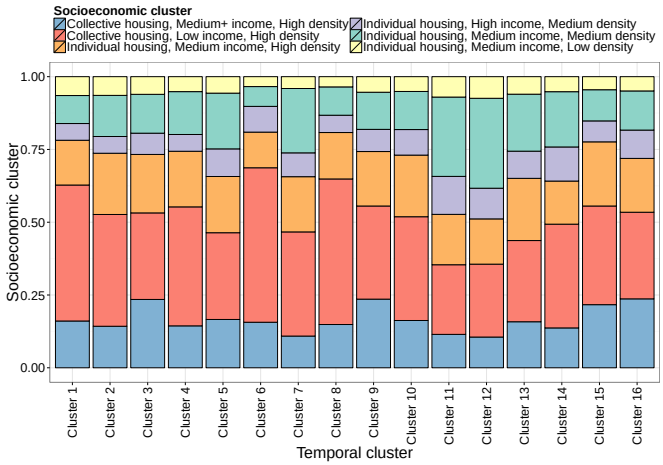


Figure 6: Proportions of socioeconomic clusters per temporal cluster.

for public transportation seems to come mainly from low income individuals and medium income individuals living in medium and high density areas. The lowest demands, on the other hand, are registered for high income individuals and medium income living in low density areas. Geographically, the latter two groups are located far from the center of Rennes, which suggests that they prefer using their private cars rather than using public transportation.

5. CONCLUSIONS

Smart card data present a unique opportunity to study passenger travel behavior in public transportation systems. In this paper, we introduced an approach to clustering passengers based on their temporal habits. By estimating a mixture of unigrams model from trips captured through smart card data, we retrieve weekly profiles (or temporal clusters) depicting different public transportation demands. By inspecting these profiles, it is possible to identify different groups of workers engaging in home-work commutes at different times of the day, students traveling to and from school, etc. We also made an attempt to link these profiles to socioeconomic data about the city in order to improve their interpretability. The results show, for example, which socioeconomic classes are more susceptible of using public transportation than the others, whether or not certain fare types are exclusively used by passengers located at particular areas of the city, etc.

Several improvements can be made based on the work presented herein. For example, the residential areas of the passengers are roughly estimated through the assignment to the most frequent first-departure station. Alternative, more robust approaches need to be investigated in order to enhance the accuracy of the pairing between smart card and socioeconomic data, and insure the quality of the interpretation of the results. We chose to cluster the passengers based solely on the boarding time of the trips they made. It would be interesting to include the spatial dimension (i.e. boarding locations) either during the clustering process or during the interpretation step (e.g. to identify the stations that are impacted by a given type of demand). The number of clusters discovered using our approach can be overwhelming to analyse. This issue can be addressed by trying to regroupe similar clusters and aggregating them into a hierarchy that is more suitable for multi-level exploration (i.e. start with a small number of coarse clusters to quickly understand the macro structure of passenger behavior, then expand interesting clusters to reveal more refined patterns). Also, we focused only on smart card data involving trips made by bus and subway. It would be interesting to study how these modes compare to and complement other transportation modes such as Bike Sharing Systems (BSS), etc. Finally, we adopted an exploratory, data-driven approach for this study. Consequently, our findings need to be examined by experts (urbanists and public transportation specialists) in order to confirm their adherence to reality.

6. ACKNOWLEDGMENTS

This work is undertaken as part of the Mobilletic project and is funded through the PREDIT (Programme de recherche et d'innovation dans les transports terrestres) program. The authors would like to thank Keolis Rennes who generously provided data for this study, especially Mr. Benjamin Bertelle for his active help and support.

7. REFERENCES

[1] B. Agard, C. Morency, and M. Trépanier. Mining public transport user behaviour from smart card data. In *12th IFAC Symposium on Information Control Problems in Manufacturing (INCOM)*, 5 2006.

[2] M. Bagchi and P. R. White. What role for smart-card data from bus systems? *Proceedings of the ICE - Municipal Engineer*, 157:39–46(7), 2004.

[3] M. Bagchi and P. R. White. The potential of public transport smart card data. *Transport Policy*, 12(5):464–474, 9 2005.

[4] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012.

[5] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, 48(3):259–302, 1986.

[6] L. Birgé and P. Massart. Minimal penalties for gaussian model selection. *Probability theory and related fields*, 138(1-2):33–73, 2007.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[8] P. Bouman, E. van der Hurk, L. Kroon, T. Li, and P. Vervest. Detecting activity patterns from smart card data. In *25th Benelux Conference on Artificial Intelligence (BNAIC 2013)*, 2013.

[9] G. Celeux, F. Forbes, and N. Peyrard. EM procedures using mean field-like approximations for markov model-based image segmentation. *Pattern Recognition*, 36(1):131–144, 2003.

[10] F. Forbes and N. Peyrard. Hidden markov random field model selection criteria based on mean field-like approximations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1089–1101, 2003.

[11] T. Fuse, K. Makimura, and T. Nakamura. Observation of travel behavior by ic card data and application to

transportation planning. In *Special Joint Symposium of ISPRS Commission IV and AutoCarto 2010*, 2010.

[12] N. Lathia and L. Capra. How smart is your smartcard? measuring travel behaviours, perceptions, and incentives. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 291–300, New York, NY, USA, 2011. ACM.

[13] N. Lathia, D. Quercia, and J. Crowcroft. The hidden image of the city: Sensing community well-being from urban mobility. In *Proceedings of the 10th International Conference on Pervasive Computing*, Pervasive'12, pages 91–98, Berlin, Heidelberg, 2012. Springer-Verlag.

[14] N. Lathia, C. Smith, J. Froehlich, and L. Capra. Individuals among commuters: Building personalised transport information services from fare collection systems. *Pervasive and Mobile Computing*, 9(5):643 – 664, 2013. Special issue on Pervasive Urban Applications.

[15] X.-l. Ma, Y.-J. Wu, Y.-h. Wang, F. Chen, and J.-f. Liu. Mining smart card data for transit riders' travel patterns. *Transportation Research Part C: Emerging Technologies*, 36(0):1 – 12, 2013.

[16] C. Morency, M. Trépanier, and B. Agard. Analysing the variability of transit users behaviour with smart card data. In *Intelligent Transportation Systems Conference, 2006. ITSC '06. IEEE*, pages 44–49, 9 2006.

[17] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using em. *Mach. Learn.*, 39(2-3):103–134, May 2000.

[18] J. Y. Park, D.-J. Kim, and Y. Lim. Use of Smart Card Data to Define Public Transit Use in Seoul, South Korea, 2008.

[19] M.-P. Pelletier, M. Trépanier, and C. Morency. Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19(4):557–568, 2011.

[20] A. Strehl, E. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In *In Workshop on Artificial Intelligence for Web Search (AAAI 2000*, pages 58–64. AAAI, 2000.

[21] W. Tran. Analysis of the differences in travel behaviour between pay as you go and season ticket holders using smart card data. In *1st Civil and Environmental Engineering Student Conference*, 6 2012.

[22] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: Concepts, methodologies, and applications. *ACM Transaction on Intelligent Systems and Technology*, 2014.