Factorisation matricielle sous contraintes pour l'analyse des usages du métro parisien

Mickaël Poussevin, Nicolas Baskiotis, Vincent Guigue, Patrick Gallinari

Factorisation matricielle sous contraintes pour l'analyse des usages du métro parisien

Mickaël Poussevin[1,2], Nicolas Baskiotis[1,2], Vincent Guigue[1,2] et Patrick Gallinari[1,2]

[1]Université Pierre et Marie Curie, PRES Sorbonne-Universités
[2]Laboratoire d'Informatique de Paris 6, UMR 7606, CNRS

Abstract

La compréhension des comportements de mobilité urbaine reste aujourd'hui pour les villes, les transports publics et les autorités de régulation une question importante et complexe qui lie géographie, urbanisme et sciences sociales. L'avènement de capteurs enregistrant les déplacements humains, de plus en plus performants, issus des cartes de transports, des capteurs routiers et des GPS mais aussi des caméras de surveillance et des réseaux de téléphonie mobile ont conduit à une explosion de la quantité de données sur la mobilité des citadins. La fouille de ces journaux d'activités pour en extraire des comportements type est une tâche difficile en raison du volume des données à traiter et du bruit inhérent à la variabilité des comportements individuels. Dans cette étude, nous proposons une approche robuste qui s'appuie sur les journaux de validations des cartes d'accès dans les transports en commun et qui apprend un ensemble d'activités latentes représentatif des usages du réseau. Nous nous concentrons sur les utilisateurs du métro parisien avec un échantillon de 600000 porteurs de cartes mensuelles ayant réalisé plus de 80 millions de trajets. Nous utilisons une représentation à double échelle temporelle des validations de chaque utilisateur. Nous extrayons de ces profils un dictionnaire d'activités latentes que nous utilisons pour caractériser les comportements individuels mais aussi les fréquentations des stations. L'analyse d'une segmentation des stations montre que les usages dans les transports publics sont liés à des facteurs géographiques et sociaux.

Mots clés: Factorisation matricielle, clustering, comportement

1 Introduction

La littérature sur la mobilité urbaine est vaste et hétérogène. Elle se compose majoritairement d'études statistiques des comportements globaux. Depuis une dizaine d'années, les études quantitatives se multiplient parallèlement au déploiement de moyens technologiques permettant de suivre les usagers: les réseaux cellulaires ont permis d'analyser les échelles des déplacements en fonction de leurs fréquences [BHG06] et plusieurs études démontrent même qu'il est possible de prévoir une grande majorité des déplacements que nous faisons quotidiennement [SQBB10]. Concernant les transports en commun, certaines études se focalisent sur des données de sondage pour caractériser les changements de comportement liés à la mise en service de nouvelles lignes [Gol02]. Les données quantitatives sont récentes et liées à l'adoption de systèmes de cartes de transport pour les usagers à Londres, Lisbonne ou encore Paris. Elles ont jusqu'ici été exploitées pour la mise en évidence des goulets d'étranglement spatio-temporels des réseaux [CSC12] ou la prédiction de certains usages comme le fait d'utiliser une ligne de bus un jour donné [FKR+13]. Cependant à ce jour, aucun article n'étudie les habitudes et usages des populations dans le temps et l'espace. Des travaux proches ont été tentés sur les taxis à Shanghai [PJW+12] ou les Vélib' parisien [RCOG13] mais sans pouvoir suivre un même utilisateur sur plusieurs trajets. Les données dont nous disposons maintenant nous permettent donc d'aborder cette tâche sous un nouveau jour en centrant notre étude sur l'usager lui-même. La mobilité urbaine est une préoccupation majeure aujourd'hui: les transports en communs sont au cœur des politiques d'aménagement du territoire et l'ANR a créé un défi spécifique sur la thématique. C'est dans ce contexte que nous proposons d'analyser les traces des usagers dans le métro parisien. Les données billettiques,

fournies par le STIF (Syndicat des Transports en Île de France), correspondent aux traces anonymisées des utilisateurs lorsqu'ils valident leur titre de transport sur les bornes du métro. Ces données sont volumineuses, 80 millions de validations et 600 000 usagers sur 91 jours, incomplètes car seule l'entrée dans le réseau est référencée et bruitées par la variance dans les activités individuelles, ce qui expliquent qu'elles n'aient pas été exploitées jusqu'à maintenant. Nous proposons ici une étude démontrant le potentiel des données billétiques pour analyser les usages du métro parisien.

Comme le montre la figure 1, nos informations sont des triplets (usager, station, temps). Les données sont riches mais difficiles à exploiter: nous proposons d'agréger les données sur trois échelles de fréquences et deux échelles temporelles. Nous distinguons donc d'une part les stations fréquemment utilisées de celles plus marginales et nous exprimons le temps de validation en fonction de l'heure de la journée et du jour de la semaine. Une fois chaque utilisateur ramené à une forme vectorielle, nous effectuons une factorisation matricielle à la manière de [PJW+12]. Nous ajoutons des contraintes de parcimonie et de forme pour obtenir une décomposition explicite: nous apprenons *in fine* un dictionnaire de fonctions temporelles correspondant à des usages. Un atome du dictionnaire représente par exemple le fait de valider à 8h45, 5 jours par semaine, ce que nous interprétons comme un départ au travail. En parallèle, chaque utilisateur est décomposé comme un sous-ensemble positivement pondéré des atomes du dictionnaire.

Cet article commence par une revue de l'état de l'art en section 2. Notre modélisation du problème et des données est présenté en section 3. L'algorithme de factorisation matricielle non-négative que nous utilisons pour extraire des comportements temporels type est présenté section 4. Nous utilisons ces comportements pour créer des profils de stations que nous étudions qualitativement en section 6. Il en ressort une corrélation entre les comportements temporels et des phénomènes géographiques et sociaux.

2 État de l'art

Nous proposons un état de l'art en deux parties qui aborde d'une part le problème de la mobilité urbaine et ensuite les techniques de factorisation matricielle et leur mode de fonctionnement.

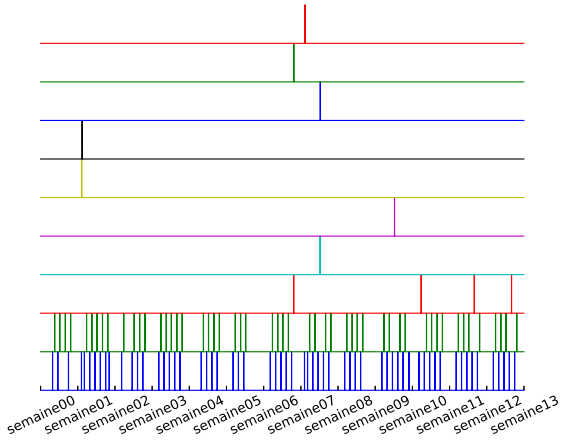

Figure 1: Validations d'un utilisateurs sur 91 jours avec ses 10 stations classées par fréquences croissante de haut en bas. Les deux lignes du bas correspondent aux stations fréquentes et probablement à la résidence et au lieu de travail.

2.1 Mobilité urbaine

Dans la littérature, le problème de la mobilité urbaine est étudié à différents niveaux. D'abord au niveau des politiques d'aménagement du territoire [BPS02] pour promouvoir des indicateurs de performance mesurables puis au niveau des déplacements eux-mêmes ensuite. Les trajets en voiture font l'objet de plusieurs études [BHG06, GHB08]: les auteurs caractérisent respectivement l'échelle des déplacements au cours du temps et les lieux de convergence fréquents d'une population cible en utilisant les réseaux de téléphonie mobile pour la localisation. L'équipe de A. Barabàsi insiste sur la prédictibilité de nos déplacements qui sont très récurrents. Dans leur étude [SQBB10], ils montrent que plus de 90% de nos mouvements sont théoriquement prévisibles. Dans la même logique, leur article suivant [WPS+11] va plus loin en mettant en parallèle les comportements de déplacements concrets et les profils des usagers dans les réseaux sociaux. Pour ce faire, les auteurs travaillent sur des données complètes de téléphonie mobile (CDR, Call Detail Record) contenant la localisation ainsi que les données échangées. Ils mettent en évidence le parallèle entre les profils virtuels et de déplacement pour conclure que ces derniers sont caractéristiques des usagers. L'étude très récente [LLC+14] propose une caractérisation des déplacements (toujours basé sur les traces collectées dans les réseaux de téléphonie mobile) pour les jours de semaine dans les 31 principales

agglomérations espagnoles. Les auteurs mettent en évidence les points chauds de chaque ville, c'est à dire les centres névralgiques rassemblant le plus de population. Ils montrent que la dynamique et la répartition de ces points est caractéristique de chaque ville.

Les données de téléphonie mobile sont aussi utilisées pour l'analyse du trafic routier et la détection d'anomalie [Her10]. [LZC+11] proposent même une analyse sur la causalité des anomalies: une première description sous forme d'arbre permet d'identifier les comportements typiques (via les branches fréquentes), puis les anomalies sont détectées comme des écarts aux chemins fréquents et l'article se concentre sur les effets de causalité entre anomalies pour isoler la source des problèmes. Les données viennent des GPS de 33000 taxis pékinois sur 6 mois et représentent 800 millions de kilomètres de trajets. [PJW+12] travaillent également sur les traces GPS de 2000 taxis de Shanghai, mais utilisent la factorisation matricielle pour mettre en évidence les habitudes collectives des chauffeurs. Ces dernières références illustrent une tendance récente: la prise en compte de différents capteurs pour localiser les utilisateurs dans leur activités. [LXMW12] décrivent l'utilisation de scanners Bluetooth pour la reconstruction des trajectoires piétonnes des visiteurs du zoo de Duisburg. Les systèmes de vélo en libre service permettent également de suivre les trajets des utilisateurs du service. Les études [BC11, RCOG13] analysent respectivement les parcours à Londres et Paris. La première référence se positionne plus sur l'analyse de graphe et la visualisation d'une grande masse de données connectées et met en avant les notions de connectivité et de centralité pour exprimer les clusters comportementaux. La seconde étude concerne 2.5 millions de trajets parisiens et permet aux auteurs de catégoriser les trajets et leur dynamique: ils analysent quelques clusters spatio-temporels centrés sur les horaires de bureau, les week-ends au parc et les transports nocturnes une fois les transports en commun fermés. Cependant, les données de cet article ne contiennent pas d'identifiant utilisateur et il n'est donc pas possible de suivre les acteurs du système. Concernant les transports en commun, plusieurs études se basent sur les analyses qualitatives et les données de sondage qui se focalisent en général sur une portion de trajet ou un usage en particulier [Gol02]. Depuis une dizaine d'année, l'adoption de systèmes d'identification des usagers à Londres, Lisbonne ou Paris ouvre la voie à des études quantitatives. [CSC12] propose une étude sur la répartition spatio-temporelle des usagers du métro londonien et isole les goulets d'étranglements du réseau dans le but de trouver des solutions pour fluidifier le trafic. L'étude porte sur une période d'un mois et se focalise sur trois stations particulières qui sont prises comme références pour modéliser les zones résidentielles, de travail et les hubs (gares). Un seuil est ensuite défini pour modéliser la surcharge. Une étude du réseau lisboète centrée sur l'utilisateur propose de personnaliser l'accès à l'information sur les problèmes de trafic dans le réseau en prédisant les usages d'une personne dans les transports [FKR+13]. Les données rassemblent 24 millions de trajets et 800 000 usagers sur 61 jours et permettent d'extraire des profils d'usage pour les jours de semaine et le week-end. Les auteurs cherchent à prédire le fait qu'une personne emprunte une ligne de bus un jour donné. Leurs conclusions rejoignent celles de [SQBB10] sur la prédictibilité des trajets fréquents, mais ils ne proposent pas d'analyse sur les comportements moins fréquents.

Notre approche est basée sur les usagers: les abonnements étant nominatifs, il devient possible de suivre les personnes et de décrire les habitudes pour extraire des usages classiques à l'échelle de la journée et de la semaine. Nous utilisons une approche à base de factorisation matricielle similaire à [PJW+12, RCOG13] mais le suivi des usagers nous permet d'ajouter la notion de périodicité dans la caractérisation des trajets.

2.2 Factorisation matricielle non-négative

Les algorithmes de factorisation matricielle sont couramment utilisés en analyse de données et parfaitement décrits dans [GVL96]. Les techniques de factorisations attaquent un problème double (apprentissage du dictionnaire et décomposition sur ce dictionnaire) qui comporte beaucoup de paramètres. Afin d'obtenir des résultats sensés, il est souvent nécessaire d'ajouter des contraintes sur la décomposition. Les plus classiques sont la non-négativité [LS00] et la parcimonie [Hoy04]. La première signifie que les éléments à décomposer (les profils d'usagers dans notre problème) sont exprimés comme une somme pondérée positivement des éléments du dictionnaire (aucune soustraction n'est admise). Il s'agit par exemple d'une hypothèse pertinente en image, où la reconnaissance de visages est formalisée comme une combinaison interprétable de caractéristiques [ZTBP06]. C'est également pertinent pour l'apprentissage d'une mixture de thèmes présents dans un corpus de documents [SBPP06] et pour l'apprentissage de densité de probabilités en général. La seconde contrainte est une forme de régularisation : le but est de reconstruire les éléments en utilisant un minimum d'atomes du dic-

tionnaire. En changeant complètement de domaine applicatif, notre approche se rapproche des travaux sur la séparation de sources (type *Independent Component Analysis*) musicale [WP05] ou de l'identification des notes dans un morceau [VBB08]. Certaines approches récentes permettent de modéliser directement les signaux composés d'impulsions [HB11], cependant la variabilité intrinsèque de nos données nous pousse plutôt vers une description lissée des données : nous considérons que la mesure de temps associée aux validations n'est pas significative à la minute près, de nombreux facteurs pouvant expliquer de légères variations.

3 Analyse et modélisation des données

Le but de notre étude est de découvrir des comportements type dans l'ensemble des journaux de validations des utilisateurs du métro et de les utiliser pour caractériser chaque trajet. Par exemple, une validation à huit heures répétée sur tous les jours ouvrés de la semaine est perçue comme un départ au travail alors qu'une validation peu fréquente le vendredi soir est associée à une sortie vespérale. Un tel étiquetage des validations caractérise à la fois l'utilisateur (son domicile, son lieu de travail et ses loisirs habituels) et les stations. Dans cette section, nous décrivons les données et leur modélisation.

3.1 Journaux de validations

Nous utilisons un jeu de données collecté par le Syndicat des Transports en Île-de-France (STIF). Plus de sept millions de personnes ont aujourd'hui souscrit à un abonnement périodique pour accéder aux transports en commun franciliens et possèdent une carte de transport dont chaque validation dans le réseau est journalisée. Ces données permettent une vue précise de l'utilisation en temps réel du réseau. L'analyse de ces journaux pour l'extraction automatique de comportements type est délicate à cause du volume (5Go/mois) qu'ils représentent mais aussi parce que l'utilisation d'un seul passager est parcimonieuse à l'échelle du système global. De plus, ils sont doublement incomplets : le STIF estime que vingt à trente pourcents des données manquent, à cause d'incident matériels ou de la fraude, et les validations ne correspondent dans la plupart des cas qu'au début du trajet, l'utilisateur ne validant pas lorsqu'il quitte le réseau.

3.2 Modélisation

Dorénavant, nous notons ce qui se réfère à l'utilisateur u, la station s et le temps t. Une validation est un triplet (u, s, t) et un utilisateur u est représenté par un journal $J_u = \{(u, s, t)\}$ composé de l'ensemble de ses validations et chacune est due à une activité latente de u. Nous proposons un modèle d'extraction de ces activités qui s'appuie sur une vue multi-échelle des données, par jour et par semaine. La première difficulté pour l'identification de ces activités latentes est la grande disparité dans la fréquence des validations (d'unique à pluriquotidien). Pour éviter que les événements rares ne soit complètement cachés par les fréquents, nous proposons de filtrer les validations en trois bandes sur un critère fréquentiel en utilisant deux seuils : la limite basse est fixée à moins d'une occurrence tous les dix jours et la limite haute à plus de deux par semaine. La bande de fréquence moyenne regroupe les intermédiaires. Nous estimons cette fréquence par utilisateur grâce à la fréquence de chaque station s dans son journal J_u. Nous considérons ensuite que la caractéristique principale d'une activité est l'heure à laquelle elle se déroule et non son emplacement géographique. Cette hypothèse nous permet d'analyser correctement les activités régulières comme le départ au travail ou le retour au domicile ainsi que les activités qui peuvent s'accomplir en des endroits différents comme aller au restaurant. Nous agrégeons donc les données par utilisateur u en fonction du temps t et indépendamment de la stations s et assimilons une validation à son heure d'occurrence. Dans une bande de fréquence données, le comportement de l'utilisateur est donc modélisé comme la probabilité $p^b(t|u)$, $b \in \{faible, moyenne, haute\}$ d'une validation de l'utilisateur. Notre objectif est d'extraire un ensemble d'activités A sur lequel décrire les validations de chaque utilisateur u : $p^b(t|u) = \sum_{a \in A} p(t|a) * p(a|u)$. Nous utilisons une représentation agrégée par jour et par semaine des validations et chaque activité a est donc également caractérisée par une probabilité d'occurrence par jour et par semaine.

3.3 Représentation des données

Le journal J_u d'un utilisateur u, composé des triplets (u, s, t) de ses validations est représenté figure 1. Pour chaque utilisateur u, nous distinguons 3 types de stations s en fonction de leurs fréquences d'utilisation: $faible$, *haute* et *moyenne* correspondant à un usage moins d'une fois tous les dix jours, plus de deux fois par semaine et entre les deux respectivement. La

séparation des bandes de fréquences permet de garantir que les événements rares ne disparaissent pas lors des approximations des techniques de reconstruction.

Nous effectuons une discrétisation temporelle par quart d'heure de la journée et par deux heures de la semaine et nous construisons une représentation vectorielle multi-échelles de $n = 180$ dimensions, $96 = 24 * 60/15$ relatives au moment de la journée et $84 = 7 * 24/2$ pour le jour de la semaine, par bande de fréquence et donc trois vecteurs par utilisateur que nous rassemblons par ligne dans trois matrices de données $\{X^{(b)} \in \mathcal{R}_+^{m_u \times n},\ b\}$ avec m_u le nombre total d'utilisateurs. La figure 2 montre un échantillon de profils utilisateurs après une telle agrégation par jour et semaine mais sans filtrage fréquentiel. Malgré le bruit relative à la variabilité des activités individuelles, cette représentation extrait des pics d'activités journaliers plus ou moins fréquents sur le semaine mais elle tend, comme débattu plus haut, à diminuer l'importance des événements rares au profit des plus fréquents.

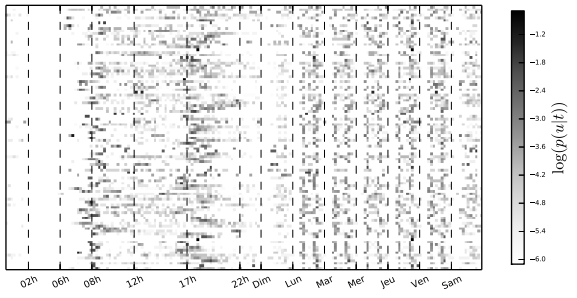

Figure 2: Profils bruts de 100 utilisateurs (un par ligne). Chaque colonne donne $\log(p(u|t))$ pour t sur la journée puis la semaine (non filtré par fréquence).

La séparation par bande fréquentielle est représentée figure 3 avec de haut en bas les bandes *haute*, *moyenne* et $faible$ et les événements relatifs aux allers-retours quotidiens entre travail et maison sont généralement présent sur la bande *haute* laissant les autres activités sur les autres bandes. Certains utilisateurs n'ont pas de validation sur une certaine bande de fréquence, ce qui en soit est déjà une caractérisation forte de leur comportement.

4 Apprentissage de comportements

Nous proposons ici un modèle robuste travaillant sur la représentation agrégée par jour et par semaine que nous venons de définir et capable d'extraire les activités latentes aux validations des utilisateurs. Nous voulons une extraction fine capable d'interpréter correctement les périodes denses comme creuses du réseau de transports publics et nous retrouvons face à un problème d'équilibre entre une représentation trop faible qui se concentrerait sur les événements quotidiens des jours ouvrés uniquement et trop descriptive proposant dans le pire des cas chaque validation comme un comportement. Nous proposons d'utiliser une factorisation matricielle non-négative pour extraire ces activités latentes. Cette technique correspond bien à notre problème car recompose chaque profil utilisateur comme une somme positive d'éléments du dictionnaire appris. Nous ajoutons de plus à l'apprentissage une contrainte de parcimonie, une contrainte sur la forme des atomes, dite mono-modale, et une normalisation.

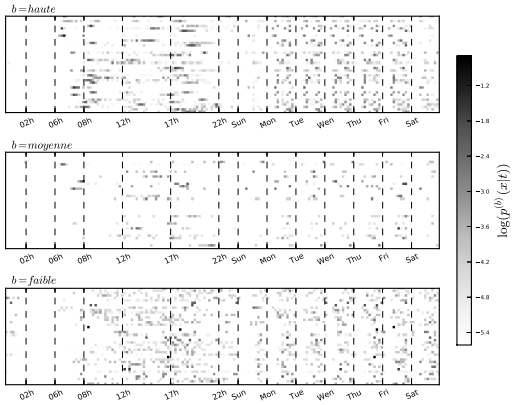

Figure 3: Profils bruts, filtrés par fréquences, de 40 utilisateurs (un par ligne). Chaque colonne donne $\log(p^{(b)}(u|t))$ filtré pour (b) valant, *haute*, *moyenne* et $faible$ (de haut en bas) et pour t sur la journée puis la semaine.

Notre objectif est d'apprendre une décomposition de chaque matrice $X^{(b)}$ (par bande de fréquence) en deux matrices positives : un dictionnaire D et un ensemble de coefficients α. Les lignes du dictionnaire D, que nous appelons atomes, représentent chacune une activité. Les lignes de α sont les coordonnées de chaque utilisateur sur dans l'espace des activités défini par D. Chaque atome du dictionnaire est représenté par une partie journalière et une hebdomadaire que nous normalisons à un. Nous imposons également que les atomes du dictionnaire soient mono-modaux, *i.e.* représentatif d'une activité particulière dans le temps. Enfin, comme chaque utilisateur ne participe qu'à un

ensemble restreint de comportements dans l'ensemble de ceux présents pour le réseau, il doit utiliser les atomes du dictionnaire avec parcimonie. Nous apprenons un dictionnaire par bande de fréquence par minimisation du coût L exprimé équation (1) dans l'ensemble des matrices non-négatives avec la contrainte C définie équation (2).

$$L(X, \alpha, D) = \frac{1}{m}\|X - \alpha.D\|^2 + \lambda|\alpha| \quad (1)$$

$$C(D) : \forall i, \sum_{j<t_{jour}} D_{ij} = \sum_{t_{jour}\leq j<t_{semaine}} D_{ij} = 1 \quad (2)$$

Ce problème d'optimisation est résolu comme présenté par l'algorithme 1 par une descente de gradient projeté sur l'espace de contraintes pour le dictionnaire D alternée avec des règles de mise à jour multiplicatives pour la matrice de coordonnées α.

$D, \alpha \leftarrow$ rand ;
tant que *convergence non atteinte* **faire**
  $D = D - \mu\alpha^T(X - \alpha D)$;
  $D = \phi(D)$;
  $\alpha = \alpha \odot \frac{XD^T}{\lambda+\alpha DD^T}$;
fin

Algorithm 1: Apprentissage de la factorisation matricielle non-négative

Pour obtenir des atomes facilement interprétables, nous utilisons régulièrement (toutes les 200 itérations) une projection additionnelle en appliquant un filtre gaussien sur la représentation de la journée de chaque atome du dictionnaire. Ce filtre, défini équation (3), où $\odot$ est le produit terme à terme, t_{pic} l'occurrence du pic et t_{jour} une variable temporelle relative à la journée, permet de ne sélectionner que le pic de plus haute amplitude pour la journée.

$$\forall i, \quad X \leftarrow X \odot \exp(-\frac{(t_{jour} - t_{pic})^2}{2\sigma^2}) \quad (3)$$

5 Analyse des représentations extraites

Nous nous concentrons sur l'utilisation du métro en excluant les bus, trains et tramways. Nous ne sélectionnons de plus que les utilisateurs abonnés mensuellement à la carte de transport et ayant suffisamment de voyages. Notre jeu de données est composé de plus de quatre-vingt millions de voyages par plus de six cent mille utilisateurs dans environ trois cents stations pendant quatre-vingt onze jours. Nous avons appris $k = 100$ atomes sur $n = 180$ dimensions avec des machines classiques (8 cœurs à 3Ghz et 16GB de RAM). Le millier d'itérations effectué par bande de fréquence prend 8 heures de calculs environ pour l'ensemble des utilisateurs. Ces atomes sont représentés figure 4.

Dans la bande de fréquence *haute*, il est intéressant de voir que la cooccurrence répétée d'une validation matinale et vespérale donne lieu à des atomes qui ne sont pas mono mais bi-modaux, malgré la projection et qui correspondent à des horaires de bureaux. De plus, les atomes sont principalement utilisés pour représenter des activités pendant les cinq jours ouvrés de la semaine et très concentrés entre huit et neuf heures. À l'opposé, les atomes du dictionnaire appris sur le bande de fréquence $faible$ sont concentrés sur la soirée et donnent une part plus importantes aux weekends. L'absence de comportement entre deux et cinq heures est dûe à l'arrêt du service sur cette plage horaire. La figure 5 montre sur la ligne supérieure, par bande de fréquence, la proportion d'utilisateurs utilisant chacun des atomes pour sa représentation, par atome et confirme que chaque atome est utilisé par une quantité similaire d'utilisateurs, assurant une utilité de chacun d'entre eux. La ligne inférieure montre, elle, l'histogramme par bande de fréquence du nombre d'atomes ayant un poids non-nul par utilisateur et montre la parcimonie dans l'utilisation des atomes du dictionnaire pour chaque utilisateur.

6 Des utilisateurs aux stations

Suite à l'extraction d'activités latentes par la factorisation matricielle non-négative, chaque utilisateur est désormais représenté comme une un vecteur pondéré sur cet ensemble d'activités. Nous souhaitons maintenant étudier la répartition de ces comportements type dans le réseau du métro. Formellement, nous voulons estimer $p(a|s)$ la probabilité d'un comportement conditionné par une station s sur l'ensemble des stations et utilisons la décomposition suivante : $p(a|s) = \sum_u p(a|u)p(u|s)$. Pour simplifier l'estimation de $p(u|s)$, nous la considérons uniforme par bande de fréquence. Nous construisons les représentations vectorielles de chaque station comme la somme par bande de fréquence des représentations vectorielles de chaque utilisateur pondérée par l'activité globale et obtenons donc trois vecteurs par station s.

Nous utilisons un algorithme de clustering multi-instance similaire à [ZZ09]. Les dimensions du problème étant faibles, 300 stations et trois fois 100 di-

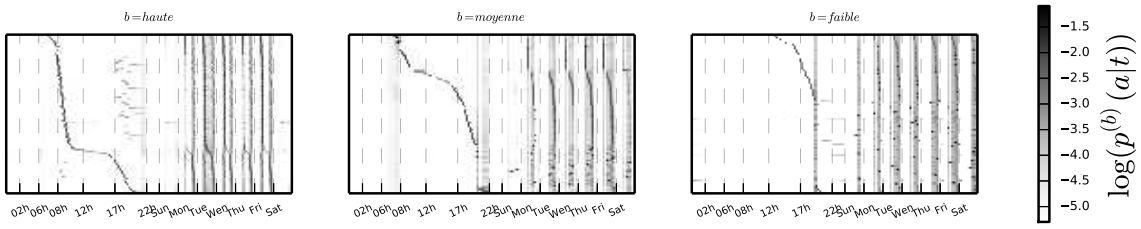

Figure 4: Atomes du dictionnaire appris par bande de fréquence. Chaque ligne est un atome, chaque colonne représente $\log(p^{(b)}(a|t))$ pour (b) valant, de gauche à droite, *haute*, *moyenne* et $faible$ et pour t sur la journée puis la semaine.

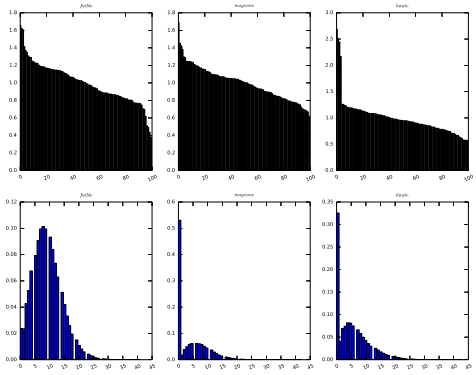

Figure 5: La ligne supérieure représente le pourcentage d'utilisation par les usagers du métro de chacun des atomes. La ligne inférieure contient les histogrammes du nombre d'atomes utilisés par utilisateur. Par bande de fréquence à chaque fois.

mensions par station, la segmentation est très rapide. Elle est également stable en fonction de l'initialisation. Nous avons extrait cinq groupes représentés sur une carte figure 6 où chaque station est colorée en fonction de son allocation par le clustering. Une répartition géographique des segments est apparente : deux groupes de stations au centre de Paris, un groupe périphérique à la limite de la ville et de sa banlieue et une séparation entre l'est et l'ouest de la petite couronne. Les deux groupes du centre séparent les quartiers touristiques de Paris des quartiers plus résidentiels et des gares. Les quartiers qui longent les portes et le périphérique entre Paris et sa proche banlieue appartiennent à un même cluster d'utilisation. Enfin, il est intéressant de voir que les banlieues ouest (Neuilly, Boulogne et Levallois) sont regroupées avec Vincennes et opposées aux banlieues est, nord et sud (Gennevilliers, Saint Denis, Montreuil, Villejuif).

Les centroïdes obtenus par la segmentation sont représentés figure 7. Chaque ligne représente un centroïde, les couleurs étant les mêmes que pour la carte, et les colonnes correspondent à l'écart à la moyenne du profil temporel, reconstruit du profil latent, de chacun des centroïdes et par bande de fréquence. Dans cette représentation un pic correspond à plus validation dans le cluster que sur l'ensemble du réseau.

À l'inverse, un trou est un manque de validations. Le filtrage par bande de fréquence et le fait que seules les entrées dans le métro sont enregistrées signifie que l'interprétation est différente par bande de fréquence. La bande *haute* correspond aux comportement fréquents des usagers donc qui *a priori* résident ou travaillent à côté d'une des stations du cluster. À l'inverse pour les deux autres bandes, ils s'agit d'activités réalisées par des personnes qui, *a priori* toujours, n'habitent pas ici mais sont venues régulièrement (bande *moyenne*) ou accidentellement ($faible$). La première observation est que le cluster périphérique correspond fidèlement à l'usage moyen du réseau, sur l'ensemble des bandes de fréquences. Les deux bandes qui captent le plus d'activités rares sont correspondent au centre de Paris qui contient en effet la majorité des musées, cinémas, bars et restaurants. Il est intéressant de noter que les deux se caractérisent par un déficit de validations le matin et que le groupe contenant les gares s'oppose à celui du centre touristique par la présence d'un pic au milieu de ce déficit matinal. Il s'agit des personnes arrivant en train transiliens aux gares, depuis les banlieues plus éloignées, et rejoignant le métro ici. Et notre modèle est capable d'extraire des activités suffisamment finement pour distinguer ces comportements. À l'opposé, les clusters des banlieues ouest et sont caractérisés par un pic d'activité le matin et le soir dans la bande de fréquence *haute* et une absence de trafic le soir dans la bande faible $faible$.

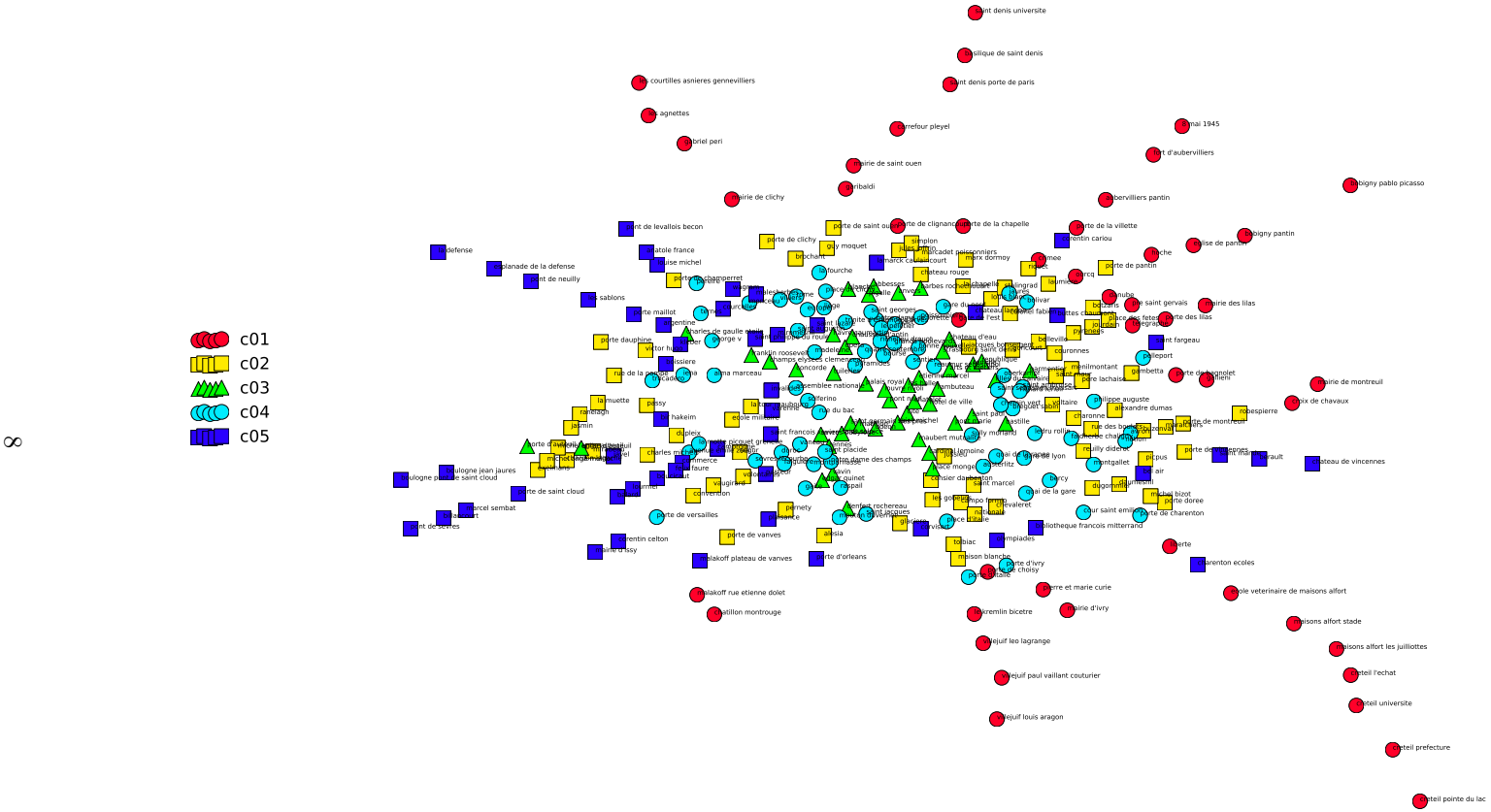

Figure 6: Carte de des stations du métro de Paris avec une coloration par cluster.

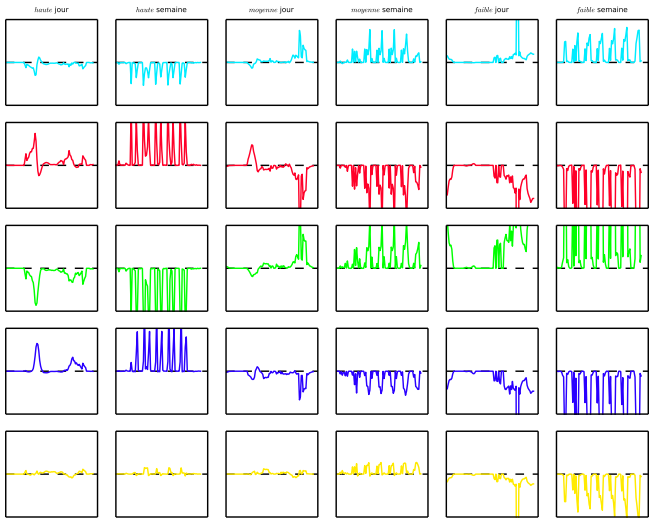

Figure 7: Différence au trafic moyen du profil des centroïdes des clusters de stations. Juxtaposition des agrégations par jour et semaine par bande de fréquence.

Ce comportement semble correspondre à ces régions résidentielles. Là aussi, la finesse de l'extraction de notre modèle sépare un départ plus tardif de la banlieue ouest d'un départ anticipé à l'est, suivi d'un déficit. Ces régions semblent donc marquées par des heures de bureaux différentes.

7 Conclusion

Nous avons dans cette étude proposé une approche robuste, s'appuyant sur l'apprentissage automatique, pour analyser la mobilité urbaine. Elle s'appuie sur les données billettiques disponible grâce à la mise en place de cartes de transports identifiant leur utilisateur. Nous avons réalisé une extraction d'activités latentes sur ces données représentées par une agrégation multi-échelle, par jour et par semaine, et sur plusieurs bandes de fréquences, de chaque utilisateur, exploitant pleinement cette nouvelle possibilité de suivre un utilisateur au cours de ses multiples trajets. Nous avons utilisé une factorisation matricielle non-négative avec une contrainte de parcimonie et de forme monomodale sur les comportements type latents. Des profils latents d'utilisateurs, exprimés comme combinaison d'activités, nous avons créés des profils de stations. Grâce à un clustering multi-instances nous avons analysés les groupes de stations issus de ces profils. Cette analyse montre que notre méthode est capable d'extraire finement des comportements type à partir d'une agrégation grossière et bruitée des journaux de validation et la segmentation sur ces profils fins révèle une répartition géographique et sociale intéressante des habitudes dans le métro Parisien. Du point de vue de l'apprentissage automatique, ces profils latents sont une représentation stable et haut niveau de données brutes bruitées et peuvent être exploitées pour mieux caractériser les utilisateurs et les stations.

Remerciements

Cette recherche a été financée en partie par le laboratoire CLEAR (Thales-LIP6) et par le projet FUI AMMICO. Nous remercions le STIF pour leurs données et leurs conseils.

References

[BC11] Anil Bawa-Cavia. Statistical analysis of dynamic urban networks. Technical report, UCL, 2011.

[BHG06] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[BPS02] John A Black, Antonio Paez, and Putu A Suthanaya. Sustainable urban transportation: performance indicators and some analytical approaches. *Journal of urban planning and development*, 128(4):184–209, 2002.

[CSC12] Irina Ceapa, Chris Smith, and Licia Capra. Avoiding the crowds: understanding tube station congestion patterns from trip data. In *ACM SIGKDD 2012*, pages 134–141. ACM, 2012.

[FKR+13] Stefan Foell, Gerd Kortuem, Reza Rawassizadeh, Santi Phithakkitnukoon, Marco Veloso, and Carlos Bento. Mining temporal patterns of transport behaviour for predicting future transport usage. In *UbiComp 13*, pages 1239–1248. ACM, 2013.

[GHB08] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[Gol02] John C Golias. Analysis of traffic corridor impacts from the introduction of the new athens metro system. *J. Transp. Geogr.*, 10(2):91–97, 2002.

[GVL96] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, 1996.

[HB11] Chinmay Hegde and Richard G Baraniuk. Sampling and recovery of pulse streams. *IEEE T. on Signal Processing*, 59(4):1505–1517, 2011.

[Her10] Ryan Jay Herring. *Real-Time Traffic Modeling and Estimation with Streaming Probe Data using Machine Learning*. PhD thesis, University of California, Berkeley, 2010.

[Hoy04] Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *JMLR*, 5:1457–1469, 2004.

[LLC+14] Thomas Louail, Maxime Lenormand, Oliva García Cantú, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. *arXiv:1401.4540*, 2014.

[LS00] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2000.

[LXMW12] Thomas Liebig, Zhao Xu, Michael May, and Stefan Wrobel. Pedestrian quantity estimation with trajectory patterns. In *MLKDD*, pages 629–643. Springer, 2012.

[LZC+11] Wei Liu, Yu Zheng, Sanjay Chawla, Jing Yuan, and Xie Xing. Discovering spatio-temporal causal interactions in traffic data streams. In *ACM SIGKDD 2011*. ACM, 2011.

[PJW+12] Chengbin Peng, Xiaogang Jin, Ka-Chun Wong, Meixia Shi, and Pietro Liò. Collective human mobility pattern from taxi trips in urban area. *PLoS ONE*, 7, 04 2012.

[RCOG13] Andry Randriamanamihaga, Etienne Côme, Latifa Oukhellou, and Gérard Govaert. Clustering the vélib origin-destinations flows by means of poisson mixture models. In *ESANN 2013*, 2013.

[SBPP06] Farial Shahnaz, Michael W Berry, V.Paul Pauca, and Robert J Plemmons. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386, 2006.

[SQBB10] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[VBB08] Emmanuel Vincent, Nancy Bertin, and Roland Badeau. Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription. In *IEEE ICASSP 2008*, pages 109–112. IEEE, 2008.

[WP05] Beiming Wang and Mark D Plumbley. Musical audio stream separation by non-negative matrix factorization. In *Proc. DMRN summer conf*, pages 23–24, 2005.

[WPS+11] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-László Barabási. Human mobility, social ties, and link prediction. In *ACM SIGKDD 2011*, pages 1100–1108. ACM, 2011.

[ZTBP06] S. Zafeiriou, A. Tefas, I. Buciu, and I. Pitas. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. *IEEE TNN*, 17(3):683–695, May 2006.

[ZZ09] Min-Ling Zhang and Zhi-Hua Zhou. Multi-instance clustering with applications to multi-instance prediction. *Applied Intelligence*, 31(1):47–68, August 2009.