École des Ponts

ParisTech

IFSTTAR

Rapport de Stage
Analyse des données de mobilités urbaines

Emile MATHIEU

Stagiaire au département Cosys laboratoire Grettia

du 21 Avril au 18 Juillet 2014

Maîtres de stage : Latifa Oukhellou, Directrice de Recherche, Ifsttar-Cosys-Grettia et Mohamed Khalil EL MAHRSI, post-doctorant, Ifsttar-Cosys-Grettia

Remerciements

Je tiens tout d'abord à remercier ma maître de stage, Latifa Oukhellou, directrice de recherche à l'Ifsttar , pour l'aide précieuse qu'elle m'a apporté tout au long du stage. Je remercie l'Ifsttar de m'avoir accueilli pendant ces trois mois et de m'avoir donné l'opportunité, à travers ce stage, de découvrir le monde de la recherche. Je tiens aussi à remercier Etienne Côme chargé de recherche et Mohamed Khalil El Mahrsi post doctorant pour avoir mis à profit leurs connaissances et m'avoir aidé durant le stage. Je remercie également mon tuteur Ecole Nicolas Coulombel pour ses nombreux conseils et remarques dans la rédaction de ce rapport.

Présentation de l'organisme d'accueil et des maitres de stage

L'Ifsttar est un Etablissement Public à caractère Scientifique et Technologique placé sous la tutelle conjointe du ministère de l'Ecologie, du Développement Durable, des Transports et du Logement et du ministère de l'Enseignement Supérieur et de la Recherche. L'Ifsttar conduit des travaux de recherche finalisée et d'expertise dans les domaines des transports, des infrastructures, des risques naturels et de la ville pour améliorer les conditions de vie de nos concitoyens et plus largement favoriser un développement durable de nos sociétés.

Le laboratoire GRETTIA exerce son activité de recherche dans le domaine de l'offre des systèmes de transport terrestres. Il s'intéresse à tous les aspects des modes routiers et transports guidés, depuis les aspects systémiques, la modélisation et la simulation jusqu'aux aspects dynamiques des véhicules en passant par la gestion, le diagnostic et la maintenance. Le GRETTIA (Génie des Réseaux de Transports Terrestres et Informatique Avancée) contribue au développement d'une ingénierie des réseaux et systèmes de transport avec la considération des problématiques d'intégration, d'intermodalité, de fiabilité et d'analyse système. Son champ de compétence inclut le domaine routier, le transport collectif et particulièrement les transports guidés.

Latifa Oukhellou, habilitée à diriger des recherches de l'Université Paris- Est (2010) et docteur de l'Université Paris-Sud Orsay (1997), est directrice de recherche à l'Ifsttar depuis 2011 où elle dirige l'équipe Diagnostic et Maintenance du GRETTIA. Elle était auparavant maître de conférence à l'Université Paris-Est Créteil entre 1998 et 2011. Ses travaux de recherches sont essentiellement articulés autour des thèmes apprentissage statistique et reconnaissance de formes pour des problèmes de diagnostic. Actuellement, ses travaux concernent le traitement de données spatio-temporelles pour des problématiques de mobilité. Elle anime un groupe de travail au sein de l'Ifsttar sur le traitement de données pour la mobilité (anim@tic). Elle a eu en responsabilité plusieurs projets de recherche subventionnés par l'ANR ou l'UE.

Etienne Côme a obtenu un diplôme d'ingénieur en informatique ainsi qu'un master de l'Université de Technologie de Compiègne en 2005. Il a ensuite effectué une thèse à l'INRETS durant laquelle il a en particulier travaillé sur l'apprentissage semi-supervisé dans le cadre d'application de diagnostic. Ces travaux l'ont ensuite amené à effectuer un post-doc à l'Université Paris 1 dans le cadre d'un projet avec la SNECMA sur des problématiques de health monitoring de moteurs d'avion à l'aide de techniques d'apprentissage non supervisé. Depuis 2010, Etienne Côme est chargé de Recherche à l'Ifsttar – GRETTIA où il développe des

recherches sur des méthodes de data mining, de machine learning et d'analyse de graphes dans le cadre d'applications sur le diagnostic de systèmes de transport, l'analyse des données de mobilité ou de données de population.

Mohamed Khalil EL MAHRSI est diplômé de l'Ecole Nationale Supérieure d'Informatique (Tunisie). Il a également obtenu un Master Of Science en génie du logiciel et informatique décisionnelle de la même école en 2009. Par la suite, Mohamed a effectué sa thèse à Telecom ParisTech où il a exploré deux problèmes de recherche liés à l'analyse et à l'extraction de données de déplacement d'objets : échantillonnage à la volée de flux de données et clustering de données de trajectoire dans des environnements de réseau routier. Il travaille actuellement comme chercheur postdoctoral à l'IFSTTAR où il est impliqué dans le projet Mobilletic.

Abstract

Bike-Sharing is a new form of sustainable transportation, such as Barclays Cycle Hire which is London' s bicycle community sharing program. Its use is analyzed focusing mainly on data recorded from July of 2012 to May of 2013. With these smart card data, travels are recorded (such as date, time, place, etc) and can further be used for analysis. The processing of these data allows us to study mobility in the city and to identify station communities using clustering. We use two approaches. The first one is based on Latent Dirichlet Allocation, a cluster analysis, revealing detailed patterns. The second method is based on Mixture of Unigrams which is also a cluster analysis, and individual patterns are highlighted. At last, we put these data into web dynamical visualization so as to explore it more intuitively.

Keywords : bike sharing system, smart card, data mining, clustering, R

Table des matières

Table des figures

Liste des tableaux

1 Introduction

Face à la concentration et à l'augmentation de la population, les métropoles revoient leurs politiques d'aménagement du territoire et de transports. C'est ainsi que nombreuses sont les villes qui ont mis en service un système de vélo en libre-service (VLS) afin d'offrir un moyen de transport propre, alternatif et complétant l'offre classique (métro, bus, tramway,...) déjà en place. Ainsi leur qualité, leur fiabilité et complémentarité sont des sujets d'investigation cruciaux. Ces systèmes et réseaux urbains produisent des grandes quantités de données qui sont de plus en plus mises à disposition, et cela ouvre le champ des possibilités en terme de traitement et d'analyse de ces données. En effet des outils de surveillance et diagnostic peuvent alors être développés pour les exploitants, de plus ces données peuvent aussi permettre une meilleure compréhension des usages des réseaux de transports et donc d'améliorer leur planification.

Cette nouvelle forme de données peut être intelligemment exploitée afin de comprendre les usages et la demande des usagers, de faire apparaître des motifs typiques de groupes d'usagers. Ces bases de données sont de taille importante et nécessitent d'être synthétisées. A cette fin, un regroupement automatique peut être réalisé afin de faire ressortir des groupes d'usagers utilisant de façon homogène les VLS.

Le présent rapport est structuré comme suit. Un état de l'art est tout d'abord présenté afin de décrire les différentes méthodes de synthèse de l'information et pour résumer les études antérieures traitant de la mobilité. Ensuite, les jeux de données utilisés seront détaillés, puis la méthodologie que l'on mettra en oeuvre sera décrite. Nous pourrons alors traiter les données (avec les algorithmes) puis les analyser afin de mieux comprendre la manière dont les usagers utilisent les VLS. Dans l'avant dernière partie de ce rapport, nous exposerons les visualisations réalisées dans le but de représenter des données VLS de manière dynamique et intelligible. Enfin, la dernière partie dépeint mon retour d'expérience personnel sur ce stage.

2 Etat de l'art

Nous allons tout d'abord présenter différentes méthodes de classifications automatiques. Ensuite, nous présenterons un état de l'art des études traitant de l'analyse des données de mobilité afin d'avoir un aperçu des avancés dans ce domaine.

2.1 Méthodes de synthèses de l'information

Les bases de données que nous possédons contiennent une quantité importante d'information pour que celle-ci puisse être analysée telle quelle. C'est pourquoi des algorithmes de classification automatique sont nécessaires afin de synthétiser cette information et de pouvoir l'analyser ultérieurement. Ceux-ci sont nombreux et s'utilisent dans des contextes souvent différents, une liste non exhaustive est présentée ci-dessous.

ACP : Analyse en Composantes Principales

L'analyse en Composantes Principales (Jolliffe, 1986) est une méthode qui consiste à transformer des variables corrélées en nouvelles variables décorrélées les unes des autres appelées "composantes principales".

Il s'agit d'une approche à la fois géométrique (les variables étant représentées dans un nouvel espace, selon des directions d'inertie maximale) et statistique (la recherche portant sur des axes indépendants expliquant au mieux la variance des données).

L'analyse en composantes principales peut donc être résumée comme la recherche d'une base maximisant la variance projetée et ainsi minimisant l'erreur de reconstruction.

Algorithme k-means :

L'algorithme des k-means (MacQueen, 1967) est un algorithme de partitionnement de données dont le but est de diviser des observations en k clusters dans lesquels chaque observation appartient à la partition dont elle est la plus proche (au sens d'une distance). Il est ainsi nécessaire de définir une distance, qui peut être la distance euclidienne mais pas nécessairement.

Étant donné un ensemble d'observations (x_1, x_2, …, x_n), où chaque observation est un vecteur de dimension d, l'algorithme k-means de regroupement vise à partitionner les n observations dans k ensembles $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ (k ≤ n) afin de minimiser la distance entre les points à l'intérieur de chaque partition et leurs centres respectifs :

$$\arg\min_S \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2$$

Algorithme standard :

- Choisir k points qui représentent la position moyenne des partitions $m_1^{(1)}$,…,$m_k^{(1)}$ initiales (au hasard par exemple)
- Répéter jusqu'à convergence :
  - assigner chaque observation à la partition la plus proche (i.e effectuer une partition de Voronoï selon les moyennes) :
    $S_i^{(t)} = \left\{ \mathbf{x}_j : \left\| \mathbf{x}_j - \mu_i^{(t)} \right\| \leq \left\| \mathbf{x}_j - \mu_{i^*}^{(t)} \right\| \; \forall \; i^* = 1, \ldots, k \right\}$

— mettre à jour la moyenne de chaque cluster :
$\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{\mathbf{x}_j \in S_i^{(t)}} \mathbf{x}_j$

Les deux phases sont itérativement répétées jusqu'à ce qu'on atteigne un critère d'arrêt, comme un nombre maximum d'itérations ou l'absence de modifications des centroïdes. Les inconvénients de cet algorithme sont l'initialisation (généralement on choisit les K premiers centroïdes au hasard) et le choix de K qui a priori n'est pas toujours évident. Le problème d'initialisation peut être résolu (Arthur and Vassilvitskii, 2007) en choisissant les points initiaux de manière astucieuse, ce qui permet de converger vers l'optimum global et non un optimum local.

Regroupement hiérarchique

Les méthodes hiérarchiques (Ju et al., 2013) se décomposent en deux principales approches. On suppose que l'on dispose d'une mesure de dissimilarité (ou similarité) entre les individus à répartir en groupes. Généralement, on utilise la distance euclidienne comme mesure de dissimilarité entre les objets.

La première méthode est une approche agglomérative d'algorithme hiérarchique ascendant. Initialement, tous les individus forment chacun à eux seuls un cluster. Ensuite on fusionne les deux clusters les plus proches puis on réitère jusqu'à obtenir le nombre de clusters souhaité. Lorsque les clusters sont constitués de plusieurs individus, il existe différentes définitions de la dissimilarité inter-classes. On peut considérer le minimum des distances entre deux points appartenant à chacun des clusters ou à des distances moyennes par exemple.

L'autre approche au contraire, est dissociative, on part d'une seule classe contenant tous les objets et on divise récursivement la classe la plus appropriée. Le processus continue tant que le critère d'arrêt (nombre de clusters fixé a priori par exemple) n'est pas vérifié.

LDA : Allocation de Dirichlet latente

Latent Dirichlet Allocation (LDA) (D. Blei, 2003) est un modèle probabiliste génératif qui permet de décrire des collections de documents de texte mais aussi d'autres types de données discrètes. LDA fait partie d'une catégorie de modèles appelés "topic models", qui cherchent à découvrir des structures thématiques cachées dans des vastes archives de documents. Les topic models sont basés sur deux hypothèses simples. La première est que dans une grande collection de textes, il existe un certain nombre de groupes ou sources de textes. La deuxième hypothèse est que les textes provenant de sources différentes tendent à utiliser un vocabulaire différent. Ainsi, le mot "bourse" aurait une plus grande probabilité de venir d'une revue économique que d'une revue traitant de l'informatique. Ces modèles permettent de synthétiser l'information et ainsi pouvoir analyser d'importantes masses de données (qui sont inextricables en tant que telles) en les regroupant par "paquets" de données qui apparaissent régulièrement ensemble. Ceci permet d'obtenir des méthodes efficaces pour le traitement et l'organisation des documents de ces archives, telles l'organisation automatique des documents par sujet, la recherche, compréhension et analyse du texte, ou même résumer des textes.

Comme l'explique l'article de Bietti (2012), LDA est un modèle Bayesien hiérarchique à 3 couches (la Figure 1 en est une représentation graphique) : chaque document est modélisé par un mélange de topics qui génère ensuite chaque mot du document.

La Figure 1 (ChangUk, 2013) représente le modèle graphique de LDA et en donne une intuition. Commençons par expliciter les différents termes et paramètres du modèle :

- Un *mot* w est la donnée discrète, correspondant à l'indice d'un mot dans un vocabulaire fixe de taille V . On peut considérer que w est un vecteur de taille V de composantes toutes nulles sauf pour la composante i où i est l'indice du mot choisi ($w^i = 1$).
- Un document est un N-uplet de mots, $\mathbf{w} = (w_1,...,w_n)$.
- Un *corpus* est une collection de D documents, $\mathbf{D} = (\boldsymbol{w}_1,...,\boldsymbol{w}_D)$.
- Les variables $z_{d,n}$ représentent le topic choisi pour le mot $w_{d,n}$.
- Les paramètres θ_d représentent la distribution de topics du document d.
- α et η définissent les distributions *a priori* sur θ et β respectivement, où β_k décrit la distribution du topic k.

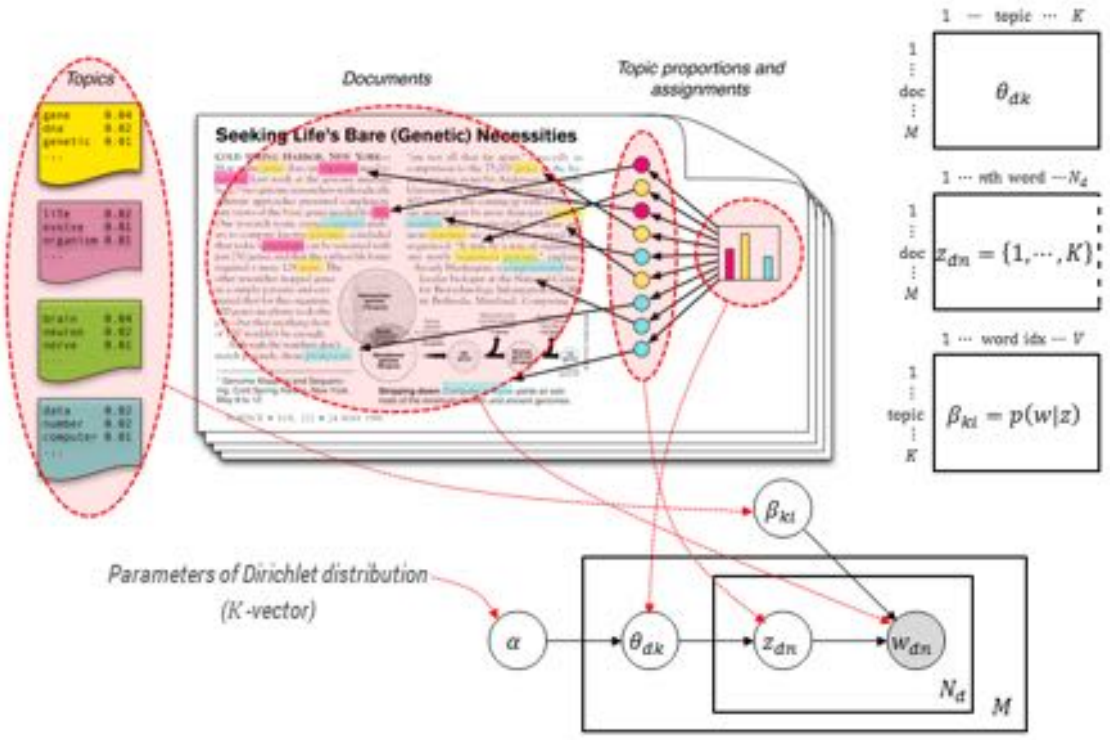

FIGURE 1 – Schéma décrivant le modèle LDA

Processus de génération Le processus génératif suivi par LDA pour un document **w** est le suivant (voir le modèle graphique de la Figure 2) :

1. Choisir $\theta \sim$ Dirichlet(α).
2. Pour chaque mot w_n :
   - Choisir un topic $z_n \sim$ Multinomial(θ)
   - Choisir un mot $w_n \sim$ Multinomial(β_k), avec k = z_n.

Cependant les variables et paramètres du modèle ne sont pas initialement connus, et il faut ainsi essayer de les apprendre à partir des données observables, c'est à dire les mots

des documents. On peut voir dans la représentation graphique de la Figure 1 que les seules variables observées sont les mots $w_{d,n}$, alors que toutes les autres variables sont cachées. Etant donnés les paramètres α et β, le rôle de l'inférence est de déterminer les variables cachées θ et z_n d'un document $\mathbf{w}$, étant donnée la liste des mots w_n du document. Les principales méthodes d'inférence (approchée) pour LDA sont les méthodes de sampling et les méthodes variationnelles.

MoU : Mixture of Unigrams

Tout comme LDA, Mixture of Unigrams (Nigam et al., 2000) est un modèle probabiliste génératif qui permet de décrire des données discrètes. Grâce à ces modèles, appartenant a la catégorie des "topic models", des structures thématiques cachées peuvent être découvertes dans ces données.

Appelons K le nombre de groupes de documents (ou topic). Chaque document (au nombre de N) est formé par n_i mots tirés d'un vocabulaire d'un nombre V de mots (uniques). Il y a trois hypothèses à considérer. Tout d'abord, nous connaissons l'origine d'un document (le topic auquel il est rattaché), ensuite nous anticiperons que les mots de ce document apparaitront probablement dans les autres documents de ce topic. Deuxièmement, puisque nous n'avons aucune information sur le document à l'avance, nous dirons qu'il est équiprobable qu'il vienne de tel ou tel topic. Finalement, puisque nous ne savons quel vocabulaire est associé avec quel topic, nous dirons qu'un mot peut apparaitre de manière équiprobable dans un document associé avec n'importe quel topic. Traduisons ceci avec des symboles mathématiques :

On tire d'abord les topics des documents :

$$z_{1:N} \overset{ind}{\sim} Multinomial(1, \theta)$$

Connaissant le topic d'un document on tire ses mots :

$$w_{ij}|z_i \overset{i.i.d.}{\sim} Multinomial(1, \phi_z i), j = 1, ..., n_i$$

Si on se place dans un cadre bayesien, on utilise les distributions a priori suivantes sur les paramètres :

$$\theta \sim Dirichlet(\alpha_{1:K})$$
$$\phi_{1:K} \overset{i.i.d.}{\sim} Dirichlet(\beta_{1:V})$$

Où w_{ij} est le jième mot du document i. z_i indique avec quel topic est associé le texte i. $\alpha_{1:K}$ et $\beta_{1:V}$ sont les paramètres pour les deux distributions de Dirichlet, ils expriment notre connaissance a priori. Est ensuite utilisé un algorithme itératif de type espérance-maximisation afin de déterminer les paramètres des distributions.

Contrairement à LDA, dans le modèle de Mixture of Unigrams, les documents ont une appartenance unique à un topic. Nous utiliserons donc le terme cluster et non topic pour ce modèle. Les usagers sont ainsi regroupés en différents clusters. Les clusters déterminés correspondent donc à des usages globaux typiques et non à des motifs plus atomiques comme pour LDA. Ainsi, les motifs bimodaux apparaissent ensemble contrairement à LDA.

2.2 Etudes antérieures sur l'analyse de données de mobilité

De nombreux travaux traitant de la mobilité et plus particulièrement analysant les données billétiques ont été réalisés. Les objectifs sont divers, ils vont de la détection des transferts multimodaux, à la détermination du taux de roulement en passant par l'étude de la variabilité des comportements et par la définition d'usagers typiques.

D'après Pelletier et al. (2011), les études existantes à propos de l'utilisation des données billettiques dans les transports publics peuvent être divisées en 3 grandes catégories :

- Les études de niveau stratégique qui tournent autour de la planification à long terme. Ces études considèrent que l'analyse des données billettiques peuvent aider à atteindre une meilleure compréhension du comportement des usagers, ce qui mène à des prévisions de la demande plus précises et contribue à prendre des décisions au long terme mieux informés.
- Les études de niveau tactique impliquant des thématiques telles que réaliser des rajustements de service, qui s'adaptent avec les grandes variations de déplacements durant les jours de la semaine et cela en faisant ressortir des usages typiques.
- Les études de niveau opérationnel, se focalisent sur la manière dont les données billettiques peuvent être utilisées afin de calculer des indicateurs de performances précis pour les transports publics (par exemple la ponctualité des horaires, le nombre de kilomètres par véhicule, le nombre de kilomètres par personne, etc). Ces études analysent aussi la détection des irrégularités et des erreurs (fraudes, équipements défectueux, erreurs humaines, etc) dans le système de paiement par smart card.

Une part importante des recherches s'intéresse à la recherche de motifs de comportements d'usagers ainsi qu'à leur variabilité. En effet Lathia and Capra (2011) discutent de la manière dont les données de paiement automatique peuvent être utilisées afin de révéler les comportements individuels et ils étudient aussi la réaction des usagers face à des mesures d'incitation. En outre Ma et al. (2013) proposent une exploration de données afin d'extraire des motifs individuels de déplacements et évaluer leur régularité, et ce avec des données incomplètes. Qui plus est, Tran (2012) analyse des différences comportementales entre les usagers du métro londonien possédant un abonnement et ceux ayant une carte "pay as you go", cela grâce aux données billettiques sur quatre mois. L'auteur propose aussi une classification des usagers en quatre catégories (visiteur, régulier, varié et divers) basée sur le nombre moyen de trajets sur chaque itinéraire et le nombre de trajets uniques sur la période couverte par les données. Ainsi, ses résultats suggèrent que les usagers n'ayant pas d'abonnement possèdent une plus grande variabilité que les autres.

Certains examinent la manière dont les trajets peuvent être liés, tels que Bouman et al. (2013) qui présentent une approche d'utilisation des données billettiques pour déduire et analyser des motifs de séquences d'activités dans le temps. Mais aussi, la détection des trajets multimodaux dans les transports publics (i.e des trajets incluant un ou plusieurs transferts entre modes de transport) grâce aux données billettiques est étudiée par Seaborn et al. (2009).

D'autres études utilisent ces données pour calculer des indicateurs de performance et de diagnostic, outils qui pourraient faciliter la gestion des réseaux de transport. En effet, Fuse et al. (2010) examinent l'utilisation des données billettiques de bus dans la détermination d'informations utiles telles que le temps de parcours et le nombre de passagers. Ces informations pourraient ensuite être utilisées pour l'analyse des points de congestion et l'amélioration de la planification des arrêts de bus. De plus, Morency et al. (2006) conduisent une étude sur la variabilité du comportement des usagers basée sur des données de validations de bus. Les auteurs mènent trois analyses afin de calculer des indicateurs globaux de performance : taux d'activité du réseau, nombre de validations par jour et l'énumération des arrêts de bus. En outre, Bagchi and White (2004, 2005) décrivent de quelle manière des "règles préétablies" [1] peuvent mesurer le taux de roulement [2], les taux de déplacements et améliorer la connaissance des trajets chaînés, et cela grâce aux données billettiques. Ils soulignent aussi le fait que puisque certaines informations ne sont pas enregistrées par les smart card, il est essentiel de continuer à récupérer des informations complémentaires de manière traditionnelle (notamment par des sondages).

Enfin Park et al. (2008) inspectent le potentiel des données billettiques pour étudier les transports publics à Séoul en Corée du Sud. En outre, les auteurs insistent sur l'importance d'examiner la fiabilité des données (c'est à dire vérifier qu'elles représentent bien l'intégralité des usagers).

Le Tableau 1 résume les objectifs, et les données utilisées des travaux existants, le Tableau 2 regroupe ces articles par thématique, et le Tableau 3 présente les données et algorithmes utilisés dans les travaux avec de la classification.

L'exploration des données billettiques dans les transports publics est une thématique qui a ainsi déjà été traitée. Cependant, comme nous allons le voir à la partie suivante, le jeu de données de Londres est riche puisque qu'il existe une colonne "client ID" qui nous permet de retracer tous les trajets des usagers.

1. Une règle peut, par exemple, déclarer que deux validations de bus utilisant la même carte en moins de 30 minutes devrait être considérer comme un unique trajet.

2. Le taux de roulement, ou turnover, définit la variation temporelle de l'utilisation d'un service, ici des transports publics.

TABLE 1 – Principaux objectifs et données utilisées dans des travaux traitant des données billettiques dans les transports publics.

Auteur(s)	Objectif(s)	Données
Bagchi and White (2004, 2005)	Détermination du taux de roulement, fréquence des trajets, et trajets chaînés.	Données billettiques des bus (Mersyside area, UK). Données billettiques des bus FiB (Bradford, UK).
Morency et al. (2006)	Analyse de la variabilité des comportements quotidiens dans les transports publics.	Trajets des bus STO (Gatineau, Canada).
Agard et al. (2006)	Analyse de la variabilité des comportements hebdomadaires dans les transports publics.	Trajets des bus STO (Gatineau, Canada).
Park et al. (2008)	Etude des variations de l'usage des transports publics.	Deux jours de données billettiques des bus et métro (Seoul, South Korea).
Seaborn et al. (2009)	Détection des transferts bus-métro, métro-bus et bus-bus.	Données des Oyster card TfL (London, UK).
Agard et al. (2009)	Définir des usagers typiques et mesurer leur comportement. Etudier la variabilité journalière, hebdomadaire et saisonnière.	Trajets des bus STO (Gatineau, Canada).
Fuse et al. (2010)	Etude des comportements des usagers en bus (durée de trajet et estimation du nombre de passagers). Analyse des points de congestion.	Données de bus (Saitama City, Japan).
Lathia and Capra (2011)	Etude des comportements des usagers (fréquence et régularité des trajets, atypicalité de motifs de déplacements et des stations visitées) et réaction aux incitations de voyage.	Données des Oyster card TfL. Résultats des codages en ligne. (London, UK)
Tran (2012)	Analyse des différences comportementales entre les usagers ayant un abonnement mensuel ou et ceux ayant une carte "pay as you go". Classification d'usagers basée sur leur comportement.	Données du métro TfL (London, UK).
Ma et al. (2013)	Extraire des motifs individuels de déplacements et étude de la régularité d'un grand et incomplet jeu de données.	Données bus et métro (Beijing, China).
Bouman et al. (2013)	Analyse temporel des motifs d'activités et de séquences d'activités.	Données billettiques OV-Chipkaart (Nederlands).

TABLE 2 – Travaux traitant de l'analyse des données billettiques par l'identification de groupes homogènes d'usagers.

Sujet	Auteur(s)
Etude comportementale des déplacements dans les transports publics	Bagchi and White (2005); Morency et al. (2006); Agard et al. (2006); Park et al. (2008); Agard et al. (2009); Fuse et al. (2010); Lathia and Capra (2011); Tran (2012).
Multimodalité et détection des trajets liés	Bagchi and White (2005); Seaborn et al. (2009).
Classification	Morency et al. (2006); Agard et al. (2006, 2009); Ma et al. (2013); Bouman et al. (2013).

TABLE 3 – Forme des données et algorithmes de classification utilisés dans la littérature traitant d'exploration de données billettiques.

Auteur(s)	Forme des données	Algorithmes de classification
Morency et al. (2006)	profils des trajets journaliers	k-means
Agard et al. (2006)	profils des trajets hebdomadaires	k-means, HAC
Agard et al. (2009)	profils des trajets quotidiens, hebdomadaires	k-means
Ma et al. (2013)	trajets chaînés, usagers	DBSCAN, k-means++ et rough set theory
Bouman et al. (2013)	durée des trajets	k-means++

3 Données et méthodologies

Dans cette partie, les jeux de données utilisés seront détaillés, tout comme la manière dont les algorithmes de classification automatiques seront mis en oeuvre.

3.1 Données

Les données sur lesquelles nous avons travaillées dans la partie 4 (Résutlats des analyses) proviennent de Transport for London (TfL), l'organisme public responsable des transports en commun de la ville de Londres. Celles-ci concernent plus particulièrement les données de Barclays Cycle Hire sur 6 mois, qui est le système de VLS de Londres [3]. Les données contiennent de nombreuses informations : Identifiant usager anonymisé, identifiant vélo, station de départ, date de départ, station d'arrivée, date d'arrivée, type abonnement. Nous pouvons noter que les informations concernant la station d'arrivée sont disponibles alors qu'elles sont souvent absentes sur les données provenant des transports en communs (type métro ou bus). On notera également que l'identifiant usager est anonymisé mais identique sur toute la durée de l'analyse, ce qui pourra nous servir à mettre en place des méthodes de recherche de typologies d'usagers, chose impossible sur les données dont dispose le laboratoire GRETTIA sur les Vélib'.

Qui plus est, dans la partie 5 - Visualisations web dynamiques, les données de Londres ont été utilisées mais aussi les données VLS de New York et Paris. Celles de New York sont disponibles en open data (données libre) sur le site de citi bike [4], contrairement aux données Vélib' de Paris disponibles au sein de l'Ifsttar dans le cadre d'une collaboration avec JCDecaux et la mairie de Paris.

3.2 Statistiques exploratoires

L'objectif est de se familiariser avec notre jeu de données londonien en réalisant des statistiques exploratoires. Sur la Figure 2, représentant la distribution du nombre de déplacements par usagers, on remarque que la majorité des usagers n'ont utilisé le VLS qu'un nombre très réduit de fois. En effet, environ 42% s'en sont servis moins de 3 fois et moins de 2% au moins une fois par jour de semaine, par conséquent une faible proportion d'usagers utilise ce système régulièrement. Afin d'obtenir des résultats pertinents sur l'analyse des groupes d'usagers, nous travaillerons par la suite uniquement sur les données de déplacements des usagers ayant fait au moins 28 trajets pendant 6 mois, ce qui correspond à au moins un trajet par semaine. La base de données est ainsi restreinte à 37 487 usagers (soit 8.56% des usagers) et 3 323 395 déplacements (soit plus de 65% des déplacements initiaux).

3. www.tfl.gov.uk/info-for/open-data-users/our-feeds
4. www.citibikenyc.com/system-data

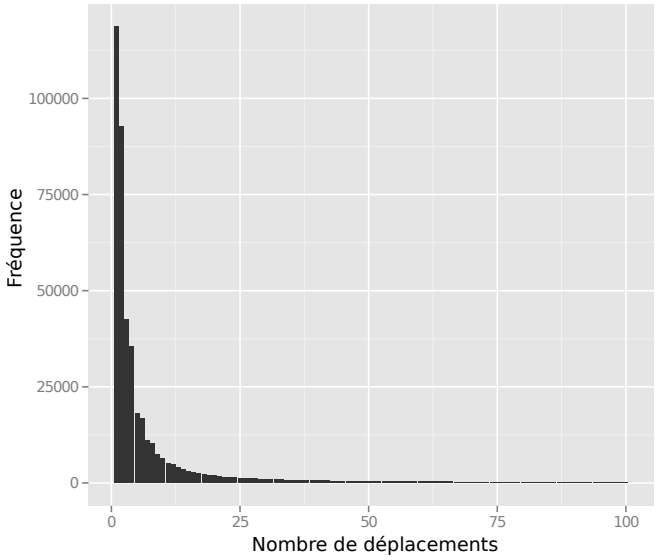

FIGURE 2 – Histogramme du nombre de déplacements par usager

Regardons maintenant la distribution sur 24h pour la semaine et le weekend pour le premier décile (des usagers ordonnés par fréquence d'utilisation sur base totale initiale), soit les usagers utilisant le système assez régulièrement. La Figure 3 illustre bien la bimodalité domicile/travail des usagers réguliers, celle-ci apparaît bien ici puisque tous les déplacements du premier décile d'usagers ont été agrégés.

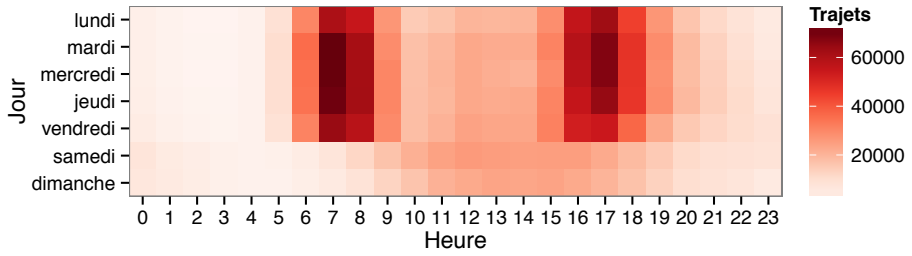

FIGURE 3 – Distribution des déplacements du premier décile sur les tranches horaires

Cette bimodalité apparaît tout aussi bien sur la Figure 4, qui est la projection de la Figure 3 sur 24h. Pour la Figure 4, les comptages sur tous les jours de semaine et du weekend étant pris en compte, puis la densité a été estimée par un noyau gaussien. On remarque d'ailleurs que la densité entre les deux pics (9 et 16h) est assez élevée et cela est dû aux trajets du weekend après-midi. La Figure 5 illustre les même résultats mais avec une distribution circulaire (puisque périodique de période 24h).

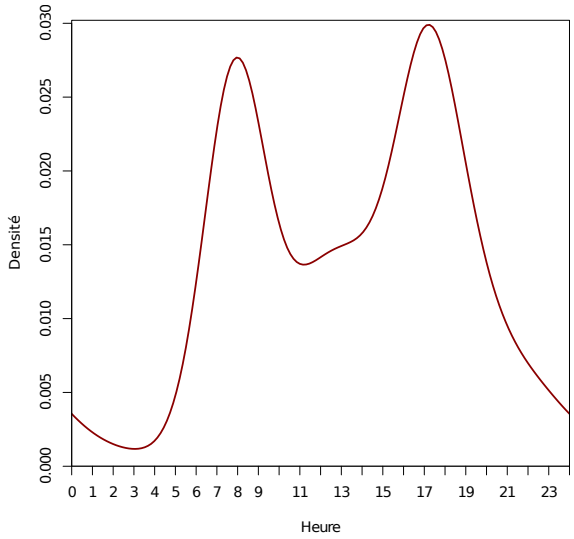

FIGURE 4 – Distribution des déplacements du premier décile sur une journée

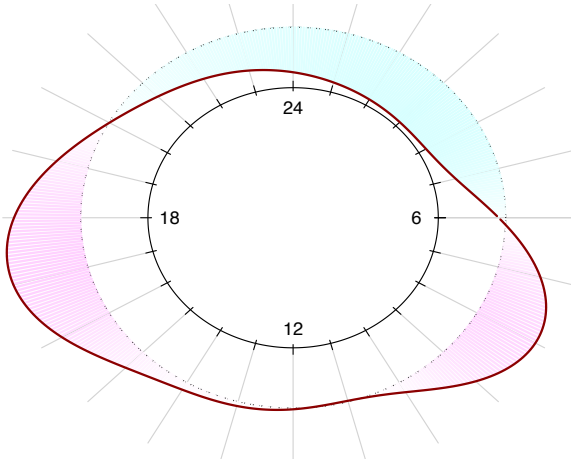

FIGURE 5 – Distribution des déplacements du premier décile sur une journée

Cependant, cette régularité n'apparaît pas chez la plupart des usagers, si l'on analyse les déplacements de trois usagers tirés au hasard dans le premier décile (Figures 6, 7 et 8), on peut en effet constater que cette régularité n'est plus visible alors que ces usagers appartiennent au premier décile et possèdent donc un certain nombre de déplacements (au moins 19 sur les 6 mois de données) à leur actif. L'usager dont les déplacements sont illustrés Figure 6 utilise uniquement le VLS en semaine et a priori principalement pour se rendre sur son lieu de travail le matin vers 8h et retourner à son domicile en milieu d'après-midi vers 16h. Cependant, un comportement plus hétérogène concernant les déplacements du weekend est observé chez l'usager de la Figure 7. Enfin l'usager dont les trajets sont présentés Figure 8 possède un comportement encore plus hétérogène, puisque ses déplacements sont très dispersés dans le temps.

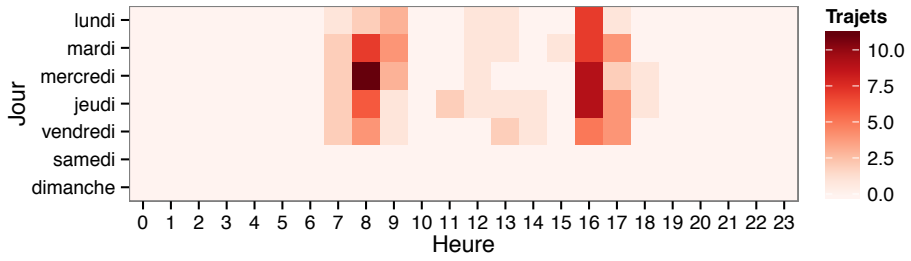

FIGURE 6 – Distribution des déplacements d'un usager du premier décile sur les tranches horaires

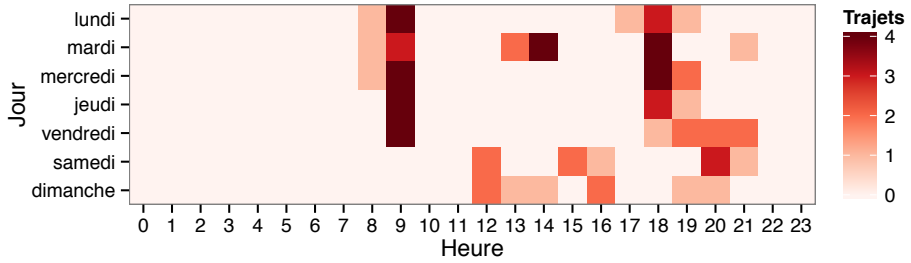

FIGURE 7 – Distribution des déplacements d'un usager du premier décile sur les tranches horaires

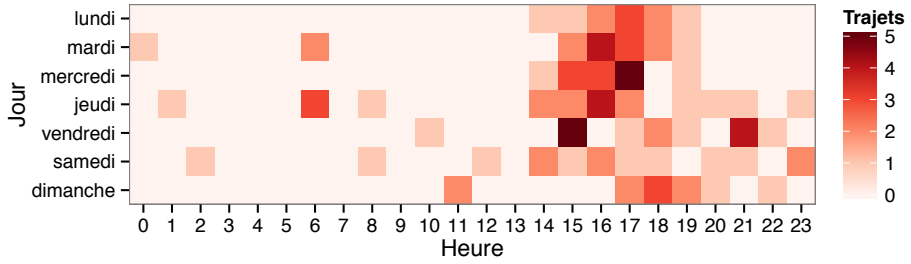

FIGURE 8 – Distribution des déplacements d'un usager du premier décile sur les tranches horaires

3.3 LDA : Allocation de Dirichlet latente

Comme expliqué à la partie 2.1 , le modèle LDA a été initialement développé puis appliqué à des collections de documents de texte. Cependant, celui-ci n'est pas limité à ce type de données et peut très bien s'utiliser dans notre cas ; il suffit de définir un ensemble de documents et de mots (vocabulaire) pertinent.

L'objectif était de faire ressortir des usages typiques des VLS de la part des usagers, et nous avons ainsi utilisé le modèle LDA pour regrouper ces usages, telles des thématiques pour un article.

En effet, LDA tout comme MoU (voir plus bas) sont spécifiques aux données de comptages, or nous possédons des variables discrètes avec des comptages sur les tranches horaires. Tandis que pour ACP, K-means et CAH, il aurait fallu définir notre espace de travail et une distance astucieuse.

Dans le modèle de notre étude, les documents sont des usagers et les mots sont des tranches horaires (du type Mardi 12-13h). En effet, un usager peut être vu comme une séquence de

trajets décrits par leur tranche horaire. Les comptages par document et mot seront alors calculés sur R. Le jeu de données est alors séparé en deux pour le futur calcul de la perplexité ; un jeu d'apprentissage et un jeu test. Nous utiliserons ensuite la fonction LDA incluse dans la bibliothèque "topicmodels" (sur R) sur le jeu de données d'apprentissage.

Nous avons lancé l'algorithme sur toute une plage de valeurs (2 à 25) pour k afin de déterminer postérieurement le nombre de clusters.

3.4 Mixture of Unigrams

Tout comme LDA, Mixture of Unigrams est un modèle probabiliste génératif qui permet de décrire des données discrètes. Grâce à ces modèles, appartenant à la catégorie des 'topic models', des structure thématiques cachées peuvent être découvertes dans ces données.

3.5 Détermination du nombre de topics/clusters dans un jeu de données

Les topic models nécessitent fixer au préalable le nombre de topics, appelons-le k, or ceci est un problème puisque celui-ci n'est pas nécessairement connu à l'avance. La détermination du nombre de clusters ou topics dans un jeu de données est toujours à l'état de recherche puisqu'il n'existe pas de méthode parfaite et que les existantes amènent souvent à différentes valeurs du nombre de clusters. Plusieurs méthodes peuvent ainsi être mise en oeuvre, tels le critère du coude, l'heuristique de la pente ou alors le recours à la perplexité.

Tout d'abord, le critère du coude consiste à détecter un coude sur le diagramme de la log-vraisemblance, et à choisir l'abscisse correspondante pour la valeur de k. En effet, de manière intuitive, plus le nombre de topics est grand, plus le gain marginal d'information diminue et donc on estime qu'à partir d'un certain nombre de topics, le gain est négligeable.

Ensuite, l'heuristique de la pente (Jean-Patrick Baudry, 2010) est basée sur le fait que des modèles n'ayant pas le même nombre de paramètres ne peuvent être comparés avec seule la vraisemblance, ainsi un critère pénalisant fonction du nombre de paramètres doit être appliqué. Un choix judicieux du coefficient pénalisant est le double de la pente de la régression linéaire appliquée à la partie pseudo-linéaire de la log-vraisemblance multiplié par le nombre de paramètres (voir Figure 9. En appliquant un tel coefficient pénalisant aux valeurs de la log-vraisemblance précédemment calculées (à la Figure 9), une courbe en cloche est alors obtenue (voir Figure 10), et la valeur maximale peut facilement être déterminée.

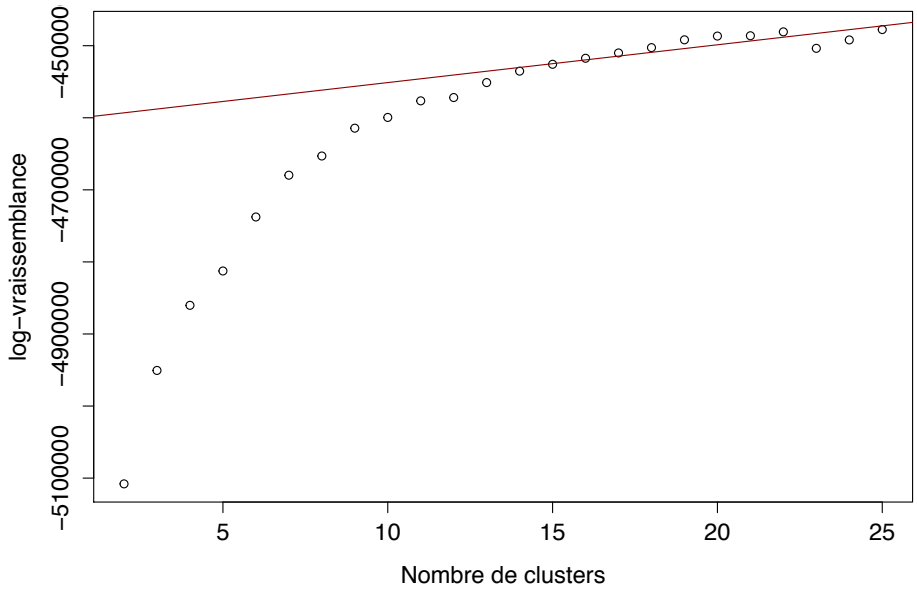

FIGURE 9 – Log-vraisemblance en fonction du nombre de topics choisi

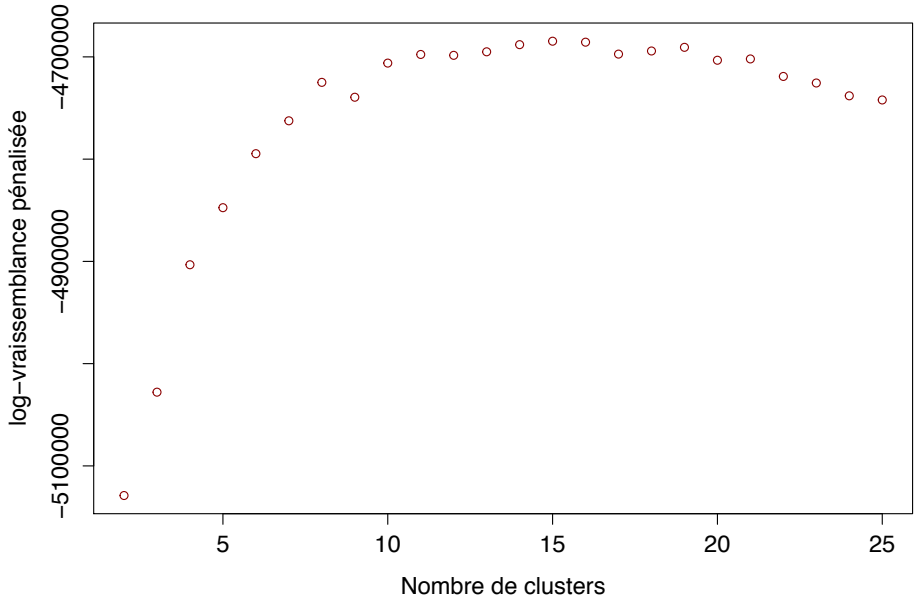

FIGURE 10 – Log-vraisemblance pénalisée, en fonction du nombre de topics choisi

Une autre manière pour déterminer ce nombre de topics n'est pas de travailler sur la log-vraisemblance mais sur la perplexité. Comme décrit dans l'article de Pleplé (2013), LDA est un cas d'utilisation classique de cette méthode, il suffit de découper le jeu de données en deux : l'un pour l'apprentissage, l'autre pour le tester. La perplexité est une transformation non linéaire de la vraisemblance calculée sur le jeu test. Ainsi, quand le modèle devient complexe (avec un nombre important de topics), la perplexité redevient croissante ; on obtient alors une courbe en cloche comme sur la Figure 11. Le nombre de topics correspond alors à l'abscisse de ce minimum.

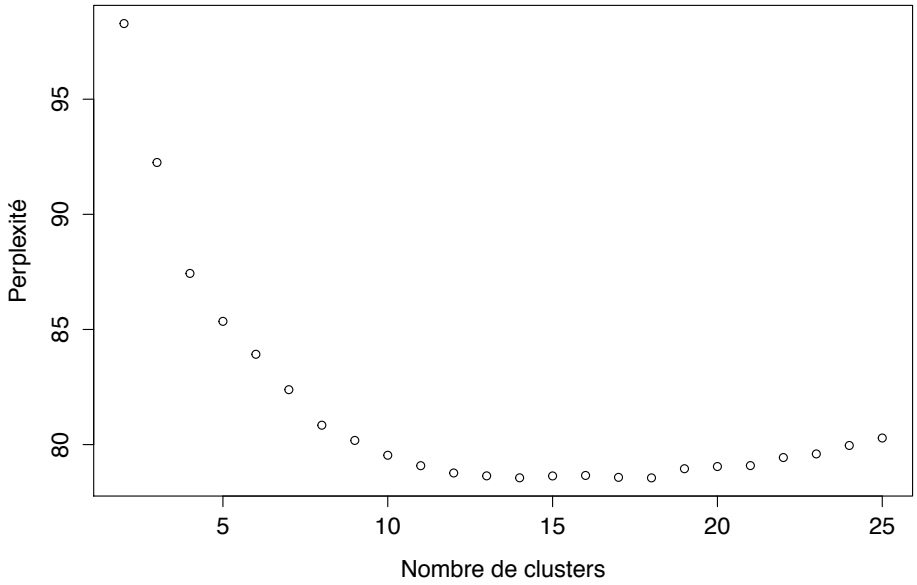

FIGURE 11 – Perplexité en fonction du nombre de topics choisi

4 Résutlats des analyses

Après avoir décrit les données sur lesquelles nous travaillons et la méthodologie employée, nous pouvons analyser les résultats des algorithmes de classification automatiques. Ceci dans l'espoir de mieux comprendre la façon dont les Londoniens se servent du système VLS.

4.1 LDA : Allocation de Dirichlet latente

Comme expliqué dans la partie 3 - Données et méthodologies, nous avons lancé l'algorithme sur une plage allant de 2 à 25 pour le nombre de topics. Comme expliqué à la partie 3.5, on utilise la perplexité pour déterminer le nombre de topics. Le minimum de la Figure 11, correspond à un nombre de topics k valant 14, c'est donc cette valeur que nous allons choisir.

Représentation des topics par des heatmap :

Ainsi nous obtenons la distribution sur les tranches horaires pour chacun de nos 14 topics ; nous pouvons penser que ces topics caractérisent assez bien le comportement des usagers car ceux-ci sont très centrés et correspondent à des usages précis. En effet, les topic 4 et 13 (Figures 23 et 22) représentent les déplacements du weekend matin et après-midi. Les topics 9, 10, 12 et 14 (Figures 21, 25, 13 et 20) illustrent les trajets de milieu d'après-midi et les trajets travail-domicile de fin d'après-midi, tout comme les topics 2, 5, 6, 7, 8 et 11 (Figures 19, 18, 26, 17, 24 et 16) illustrent ceux domicile-travail du matin. Le topic 5 (Figure 18) illustre les déplacements du déjeuner, tandis que le topic 1 (Figure 15) représente les déplacements du soir.

En outre il est intéressant de connaître la distribution agrégée (sur tous les usagers) des topics, puisque nous pouvons voir sur la Figure 12 que les topics 2 et 8 sont plus fréquents que les topics 5, 6, 11 et beaucoup plus fréquents que le topic 3 ce qui montre que le pic du matin est centré vers 7-8h. Une même analyse peut être menée pour les trajets de fin d'après-midi.

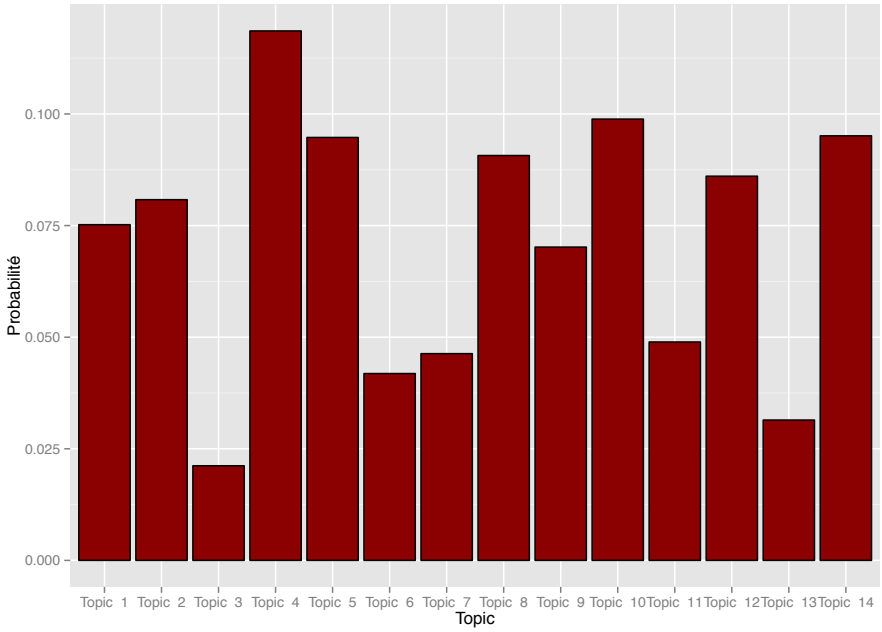

FIGURE 12 – Répartition des topics chez les usagers

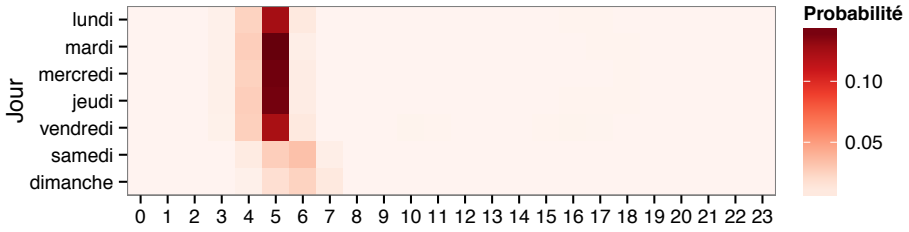

FIGURE 13 – Distribution du topic 12 sur les tranches horaires

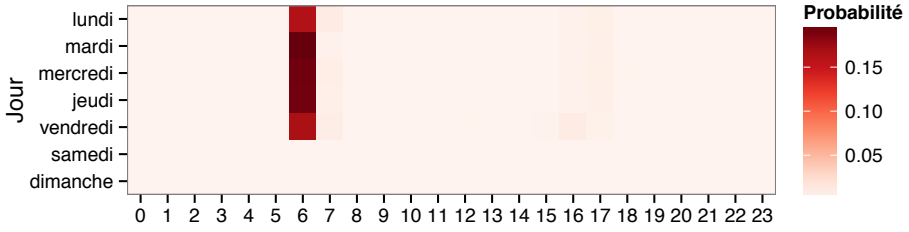

FIGURE 14 – Distribution du topic 3 sur les tranches horaires

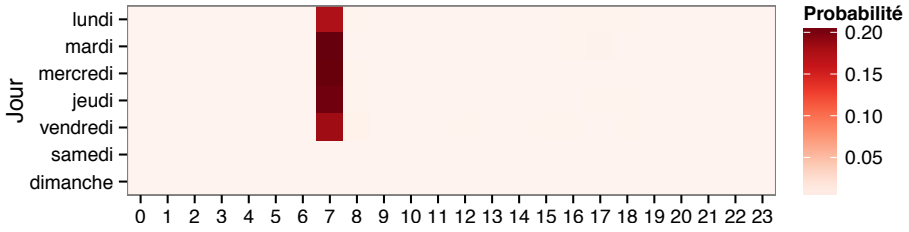

FIGURE 15 – Distribution du topic 1 sur les tranches horaires

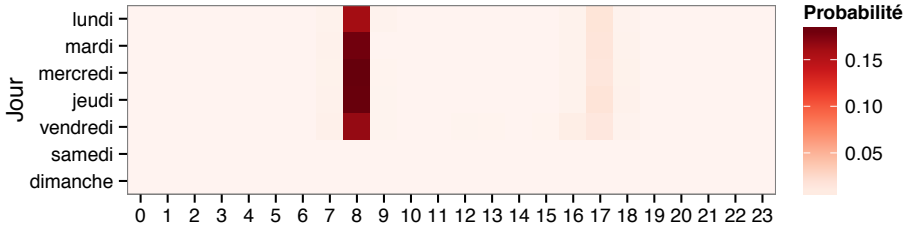

FIGURE 16 – Distribution du topic 11 sur les tranches horaires

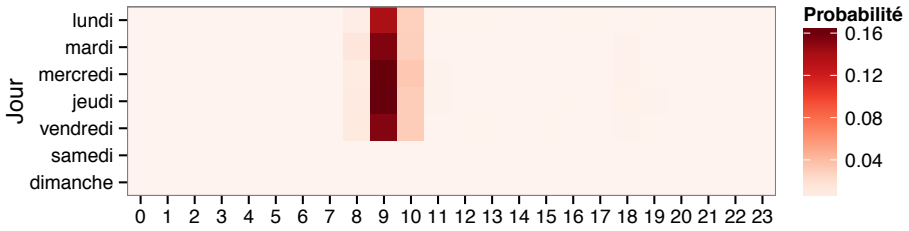

FIGURE 17 – Distribution du topic 7 sur les tranches horaires

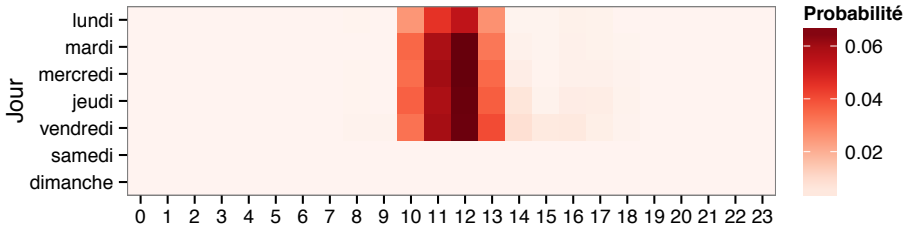

FIGURE 18 – Distribution du topic 5 sur les tranches horaires

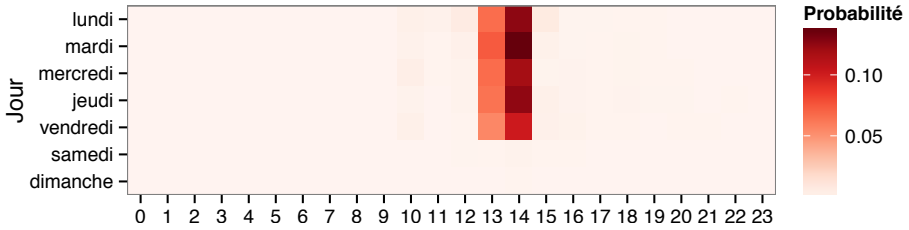

FIGURE 19 – Distribution du topic 2 sur les tranches horaires

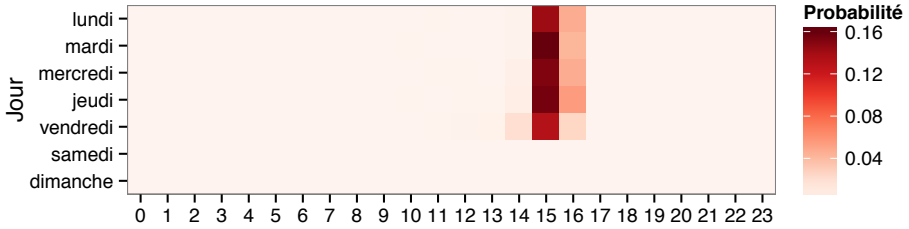

FIGURE 20 – Distribution du topic 14 sur les tranches horaires

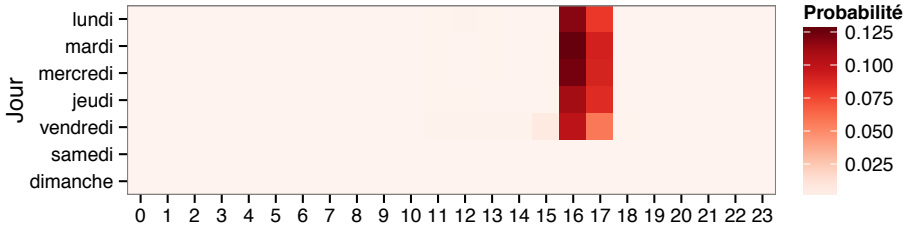

FIGURE 21 – Distribution du topic 9 sur les tranches horaires

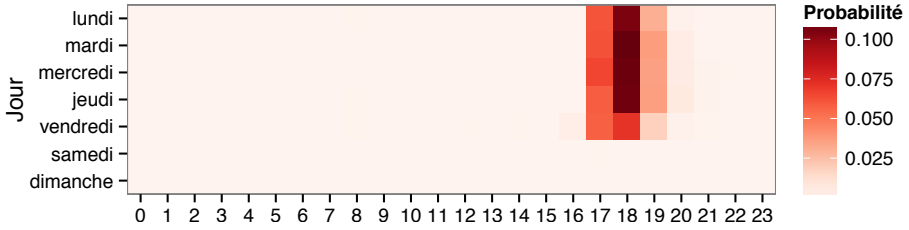

FIGURE 22 – Distribution du topic 13 sur les tranches horaires

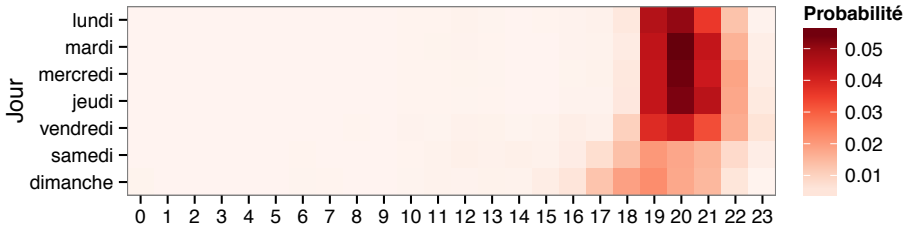

FIGURE 23 – Distribution du topic 4 sur les tranches horaires

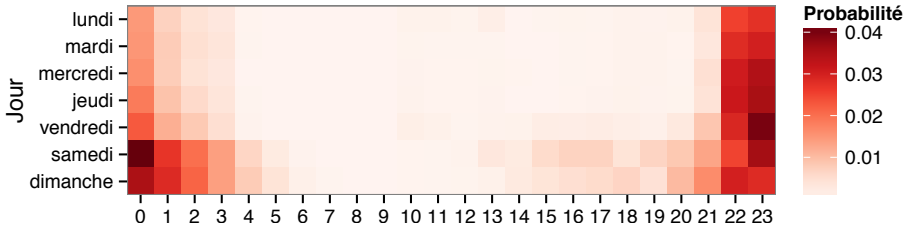

FIGURE 24 – Distribution du topic 8 sur les tranches horaires

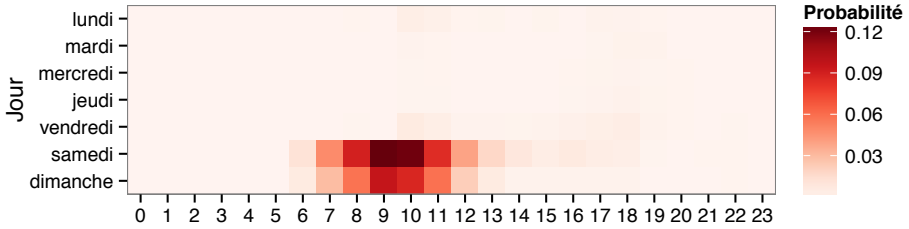

FIGURE 25 – Distribution du topic 10 sur les tranches horaires

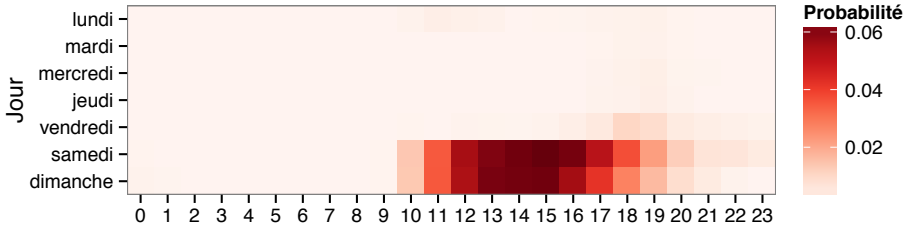

FIGURE 26 – Distribution du topic 6 sur les tranches horaires

Analyse spatiale des topics :

Nous nous sommes ensuite intéressés à la différence d'usage des stations en fonction de chaque topic, en séparant les stations d'origine et de destination des trajets. Pour se faire nous avons rassemblé dans un tableau le comptage du nombre de déplacements agrégés par tranches horaires pour chaque station, puis nous l'avons pondéré par la distribution des topics sur les tranches horaires. Il apparaît clairement sur les topics représentatifs des trajets domicile-travail (*ie.* 10, 12, 14, 2, 5, 6, 7, 8 et 11) que les zones d'habitation et de travail sont hétérogènes ; en effet on observe pour le topic 2 (Figures 27 et 28) (représentant les trajets autour de 13-14h en semaine) de nombreux départs pour les stations en périphérie et de nombreuses arrivées pour celles situées plus au centre de Londres.

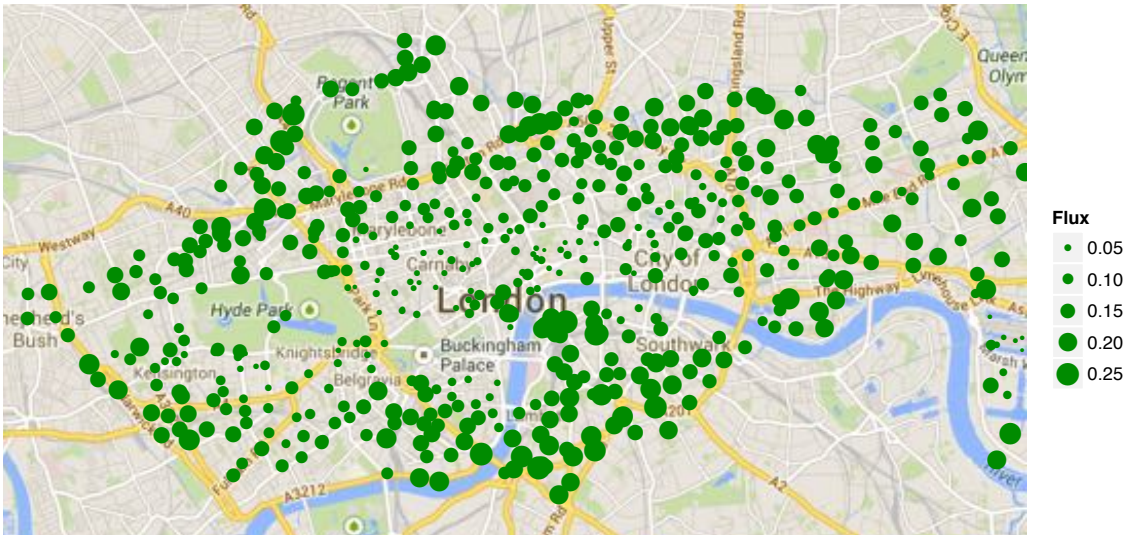

FIGURE 27 – Flux des stations d'origine pour le topic 2

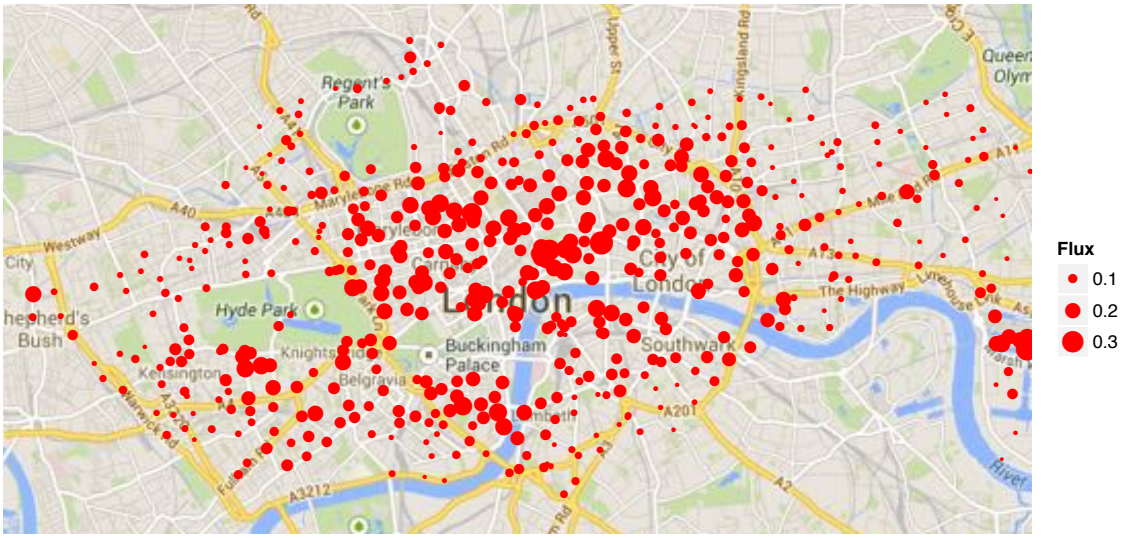

FIGURE 28 – Flux des stations de destination pour le topic 2

Qui plus est, on peut noter sur les Figures 29 et 30 que concernant les trajets du matin tôt (5h et 6h), les stations de destination des topics 5 et 11 les plus utilisées sont proches des principales gares londoniennes telles Waterloo, St Pancras, Kings Cross, Paddington, Fenchurch Street,...

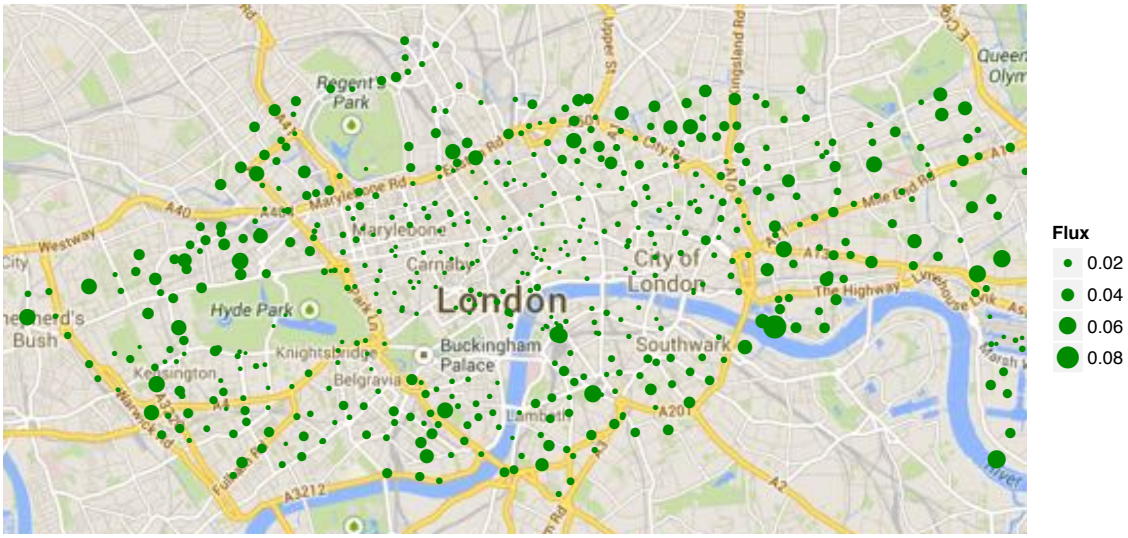

FIGURE 29 – Flux des stations d'origine pour le topic 3

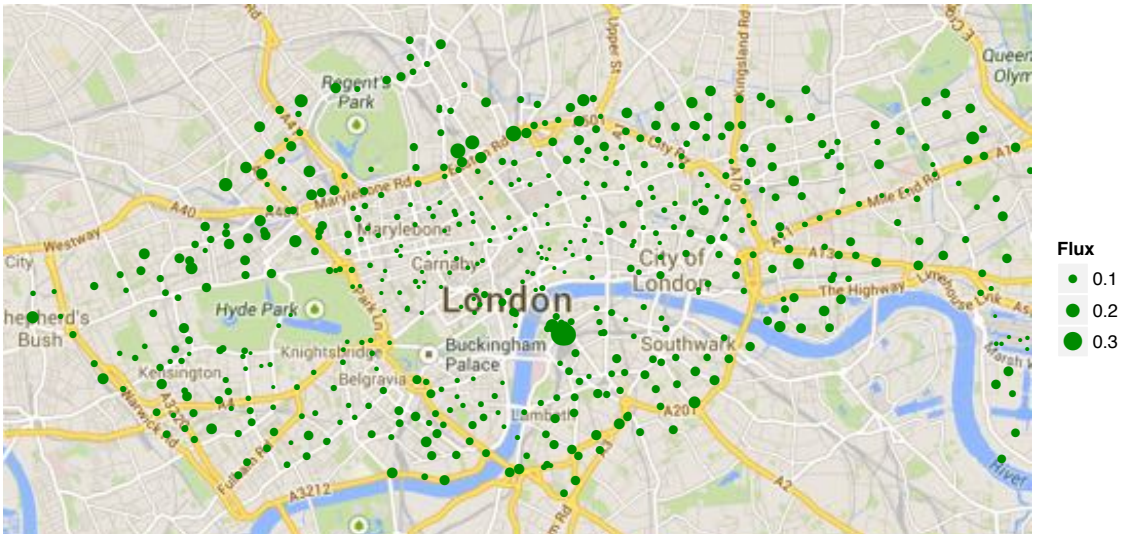

FIGURE 30 – Flux des stations d'origine pour le topic 11

Mais aussi, on peut remarquer le fait que le système ne se déséquilibre pas ou peu durant cette période, la plupart des trajets étant sans doute des boucles (retour à la station d'origine). Cela se traduit par le fait que les Figures 31 et 32 se ressemblent fortement.

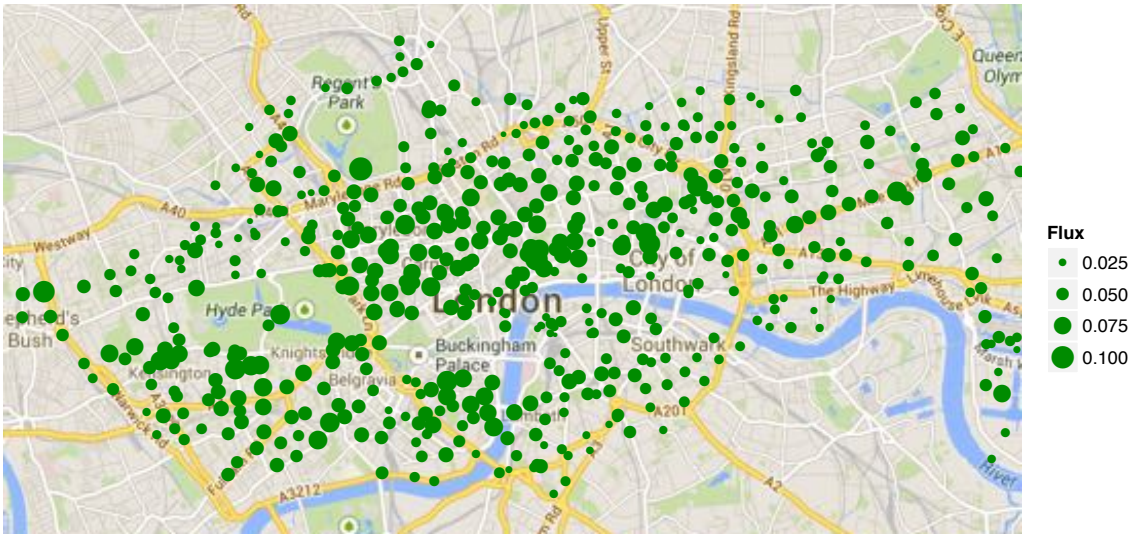

FIGURE 31 – Flux des stations d'origine pour le topic 5

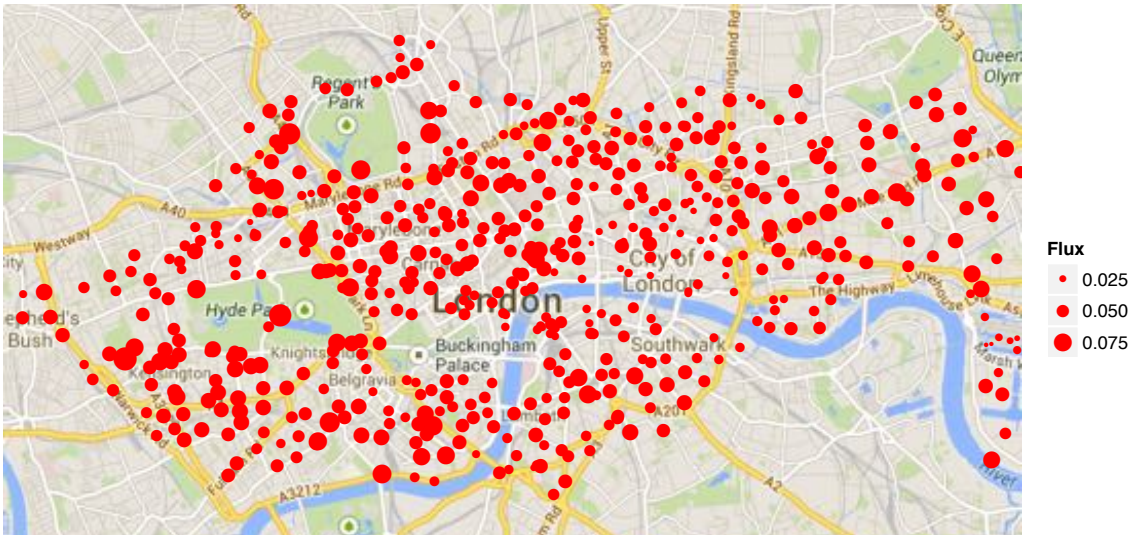

FIGURE 32 – Flux des stations de destination pour le topic 5

En outre, un autre aspect intéressant qui apparaît grâce à ces visualisations est l'utilisation spécifique des VLS les jours de weekend. En effet sur la Figure 33 qui illustre la répartition des flux de destinations du topic 4 (représentatif des déplacements du weekend), il semble que les stations les plus utilisées soient situées aux abords des parcs et espaces verts. Le système VLS est donc ici utilisé pour rejoindre des lieux de loisirs.

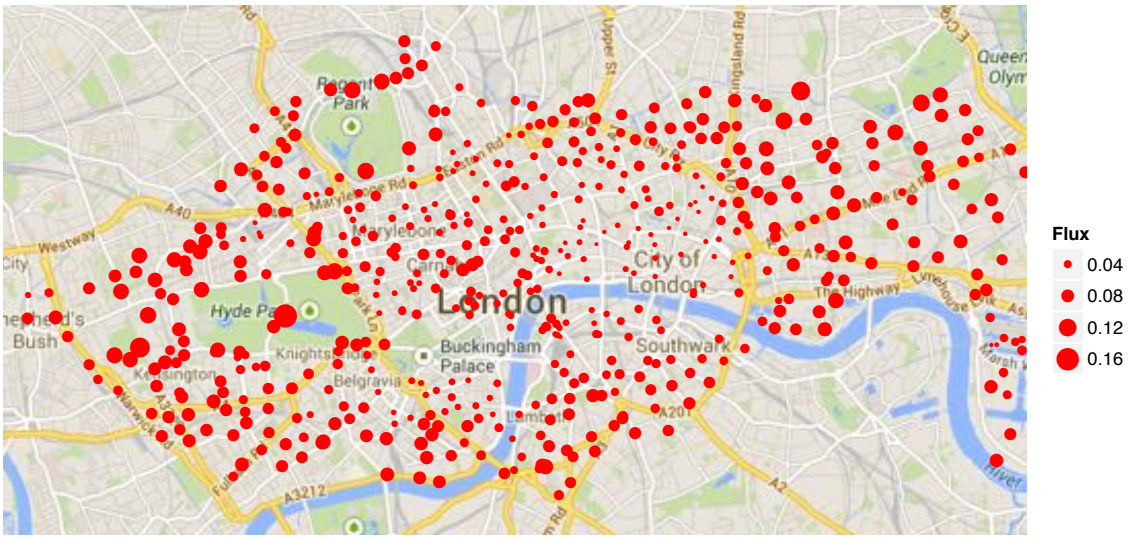

FIGURE 33 – Flux des stations de destination pour le topic 4

Corrélation entres les différents topics :

La particularité de LDA est que cet algorithme détermine des topics atomiques et que chaque usager est généralement représenté par plusieurs de ces topics. Il serait donc intéressant de savoir quels topics apparaissent fréquemment en même temps chez les usagers afin d'en déduire des usages "typique".

Nous avons ainsi déterminé la matrice de corrélation de Pearson (Benesty et al., 2009) des topics grâce à la fonction "cor" sur R avec l'option method="pearson" puis nous l'avons exportée (grâce à la bibliothèque rgexf) sur le logiciel libre d'analyse et de visualisation de réseaux Gephi, afin d'afficher graphiquement ces corrélations. Cette visualisation sur le logiciel Gephi n'est pas convaincante puisque les nombreux liens se superposent. Ainsi, nous avons simplement dessiné celle-ci grâce au package "corrplot", qui représente la matrice de corrélation avec un dégradé allant du rouge au bleu selon de coefficient de corrélation.

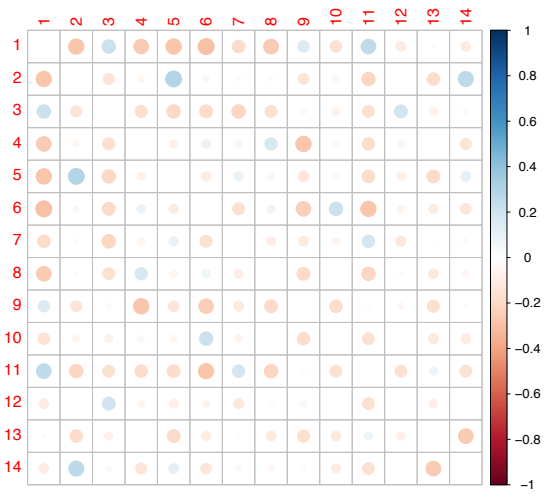

FIGURE 34 – Illustration graphique de la matrice de corrélation de Pearson des topics

En analysant la Figure 34, différents types de corrélations apparaissent. Tout d'abord, toutes les corrélations positives proviennent de topics étant temporellement proches, tels les topics 1 et 11 (7 et 8h du matin en semaine) mais aussi les topics 2 et 5 (midi et 14h en semaine) ou 2 et 14 (14 et 15h en semaine). Ceci provient probablement du fait que malgré une certaine régularité, un usager peut partir à 7h50 un matin et le lendemain à 8h10. En outre, les topics 6 et 10 (matin et après-midi le weekend) sont relativement corrélés, et ceci parait légitime puisque le même individu se déplace de son domicile à un lieu d'activité puis retourne chez lui. Cependant, les corrélations négatives sont plus intéressantes, puisque l'on peut voir que les topics 1 et 4 (6h et soir en semaine) sont négativement corrélés (il est compréhensible que les usagers se levant tôt ne sortent pas tard le soir en semaine). De manière générale, les topics représentants les trajets du matin relativement tôt en semaine (3, 1 et 11) sont négativement corrélés avec les topics représentant les trajets du soir en semaine (4 et 8).

4.2 Mixture of Unigrams

Nous avons procédé de manière analogue à LDA en lançant l'algorithme pour différentes valeurs du nombre k de topics, puis en traçant la log-vraisemblance (voir Figure 9) afin de déterminer a posteriori la valeur optimale de k.

Nous utilisons l'heuristique de la pente décrite à la partie 3.5 afin de choisir une valeur de k. Nous trouvons ainsi 14 pour le nombre k de clusters.

Contrairement à LDA, dans le modèle de Mixture of Unigrams, les usagers ont une appartenance unique à un topic. Nous utiliserons donc le terme cluster et non topic pour ce modèle. Les usagers sont ainsi regroupés en différents clusters. Les clusters déterminés correspondent donc à des usages globaux typiques et non des motifs plus atomiques comme pour LDA.

Ainsi, les motifs bimodaux apparaissent ensemble contrairement à LDA ou l'aller et le retour sont séparés. On peut ainsi remarquer que les usagers du cluster 1 (Figure 36) restent a priori plus longtemps sur leur lieu de travail que ceux du cluster 2 (Figure 42). Mais aussi que certains usages sont simplement translatés temporellement, tel le décalage d'une heure dans le comportement des usagers du 2 et 4 (Figures 42 et 45). Cependant, les motifs du weekend (Figure 39) et du soir (Figure 46) sont toujours présents. D'autres utilisent le système de manière beaucoup plus diffuse comme le montre la Figure 44. Les usagers du cluster 5 (Figure 47) eux l'utilise dans l'après midi et le soir que ce soit en semaine ou le weekend. D'autres clusters sont plus atypiques, tels les cluster 12, 4 et 9 (Figures 44, 45 et 46), puisque ceux-ci ne possèdent qu'un seul trajet dans la journée (12-17h, 16-17h et 18-19h) contrairement aux autres clusters. Qui plus est, la seule chose qui différencie les clusters 11 et 3 (Figures 38 et 39) est que les usagers du cluster 3 utilise aussi les VLS le weekend (mais dans une moindre mesure).

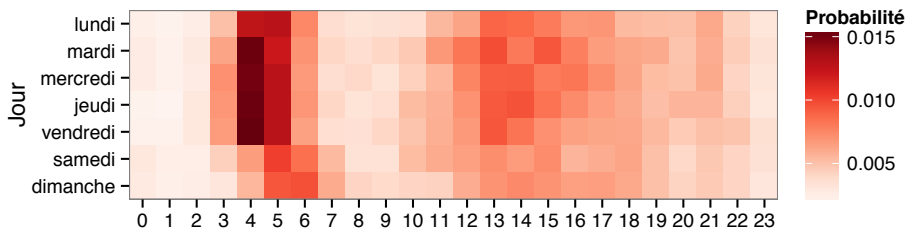

FIGURE 35 – Distribution du cluster 13 sur les tranches horaires

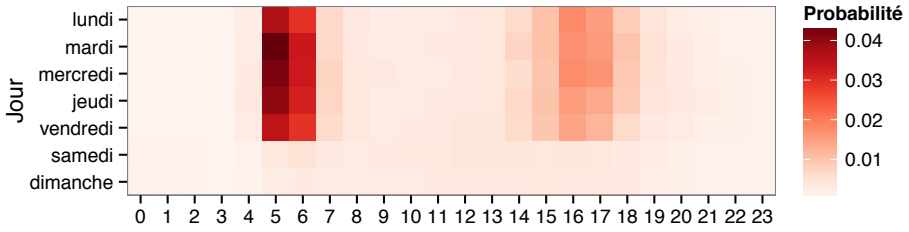

FIGURE 36 – Distribution du cluster 1 sur les tranches horaires

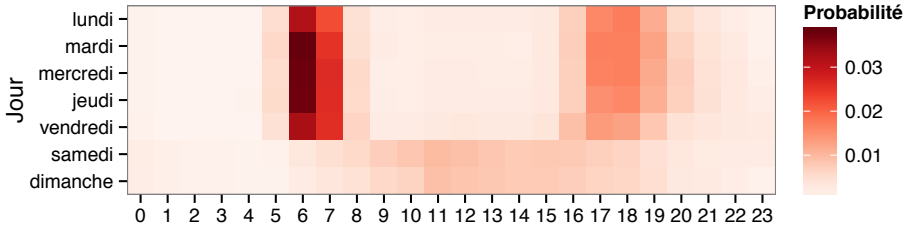

FIGURE 37 – Distribution du cluster 10 sur les tranches horaires

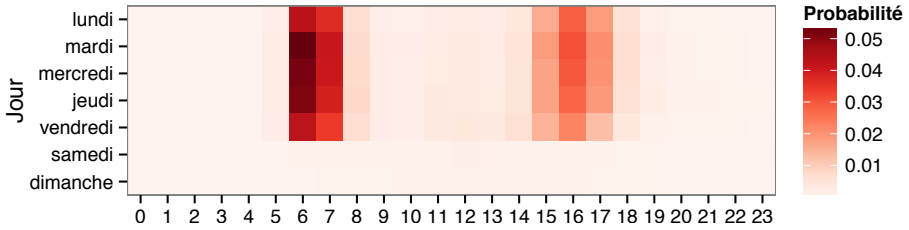

FIGURE 38 – Distribution du cluster 11 sur les tranches horaires

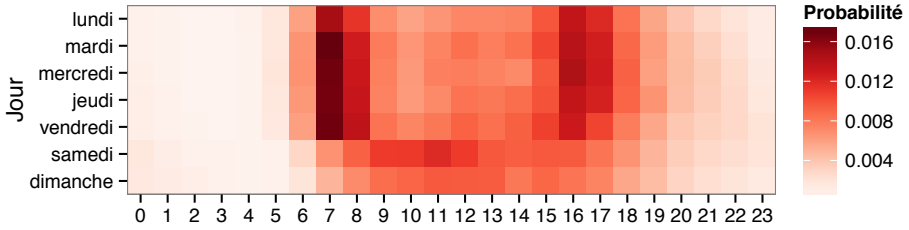

FIGURE 39 – Distribution du cluster 3 sur les tranches horaires

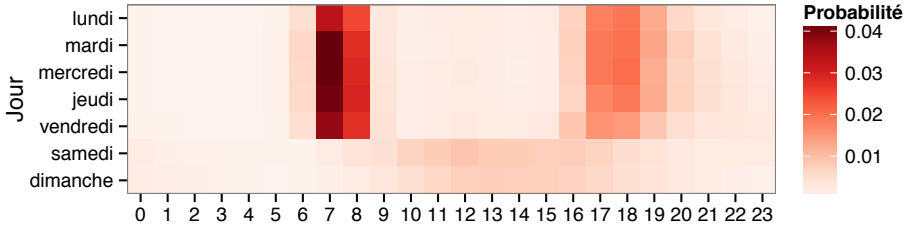

FIGURE 40 – Distribution du cluster 8 sur les tranches horaires

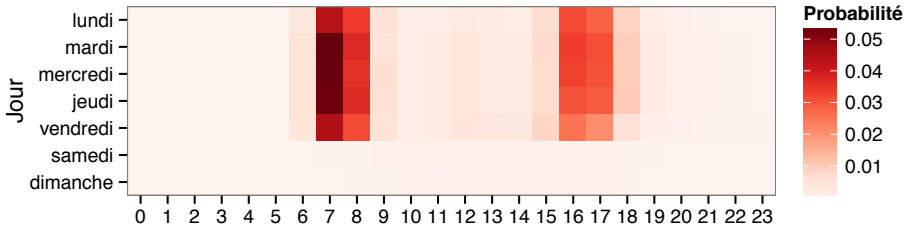

FIGURE 41 – Distribution du cluster 14 sur les tranches horaires

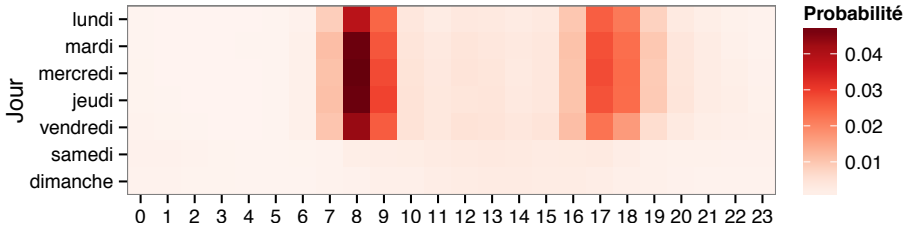

FIGURE 42 – Distribution du cluster 2 sur les tranches horaires

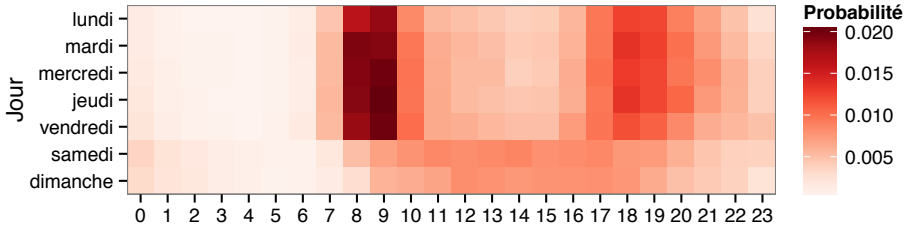

FIGURE 43 – Distribution du cluster 7 sur les tranches horaires

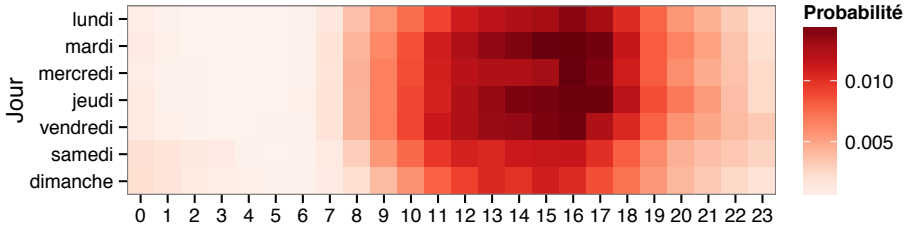

FIGURE 44 – Distribution du cluster 12 sur les tranches horaires

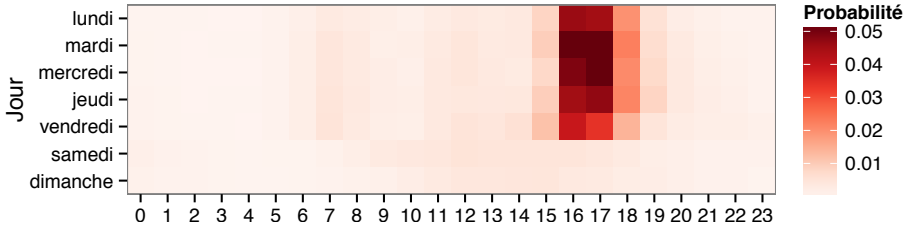

FIGURE 45 – Distribution du cluster 4 sur les tranches horaires

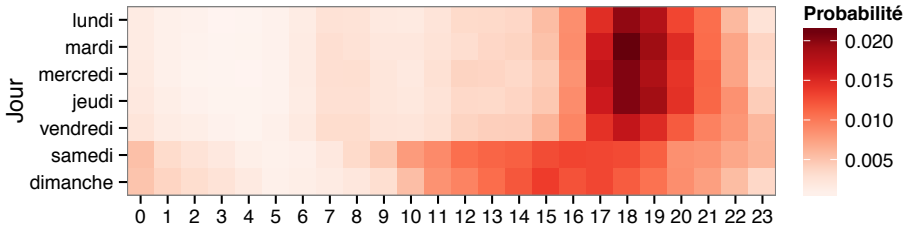

FIGURE 46 – Distribution du cluster 9 sur les tranches horaires

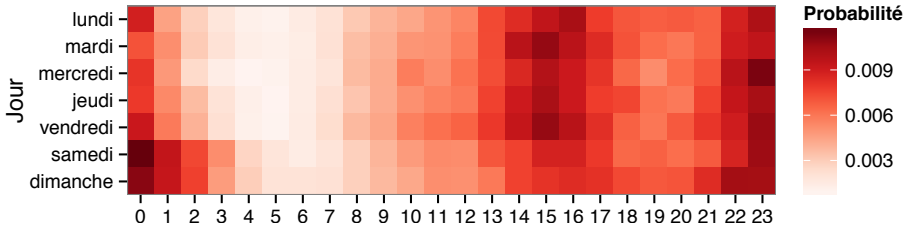

FIGURE 47 – Distribution du cluster 5 sur les tranches horaires

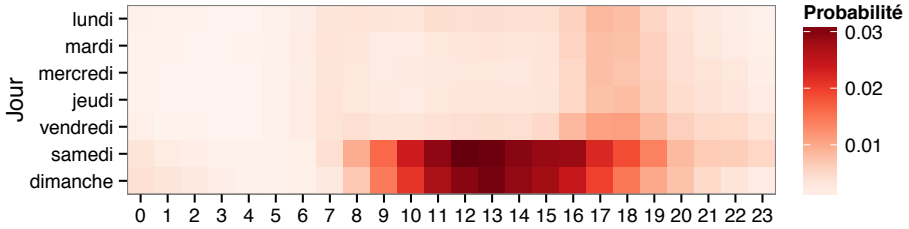

FIGURE 48 – Distribution du cluster 6 sur les tranches horaires

Les topics et clusters générés par les algorithmes de classification que sont LDA et MoU représentent différentes facettes des VLS. En effet, certains groupes l'utilisent de manière régulière et ponctuelle pour des trajets domicile-travail, alors que d'autres y ont recours plus pour un usage récréatif en particulier le soir ou les weekends.

5 Visualisations web dynamiques

Les bases de données billetiques sur lesquelles nous travaillons, comportent des millions de lignes, et il n'est pas toujours trivial de produire des visualisations claires et synthétiques. Ainsi, Etienne Côme a proposé la réalisation, durant une à deux semaines, de mini-projets de visualisations web. Utiliser les langages web, permet en effet une grande diffusabilité des travaux, puisque qu'il suffit pour y accéder d'une simple connexion internet avec un navigateur web. L'idée est de présenter de manière dynamique, grâce à une interface graphique, des données billétiques sous un certain éclairage. Une petite formation par Etienne Côme sur les langages web (HTML,CSS et Javascript) et sur des bibliothèques spécifiques (Leaflet.js, d3.js) a été nécessaire.

5.1 Projet 1 : atNight

Le premier projet sur lequel j'ai travaillé est nommé "atNight" puisque l'objectif est de visualiser l'activité nocturne de certaines métropoles grâces aux enregistrements des déplacements des VLS. Les données libres des déplacements VLS des villes de New York, Londres et Paris ont dû être téléchargées puis traitées grâce au language R. Ainsi, après avoir effectué le comptage des trajets entre 22h et 3h par stations, il a fallu exporter ces données (au format csv) pour les utiliser avec les bibliothèques Javascript. Ensuite, Leaflet.js est une bibliothèque libre de cartographie intéractive que j'ai donc utilisé comme support de base pour la visualisation. En outre, D3.js une bibliothèque d'affichage de données numériques sous forme graphique et dynamique, a permis d'ajouter un calque sur la carte avec les informations que nous voulions afficher.

Etant un néophyte du langage javascript, malgré des connaissances et de la pratique en HTML et CSS, je n'ai pas essayé de me lancer tête baissée dans le code. Je suis reparti d'une visualisation d'Etienne Côme, où celui-ci avait implémenté du code javascript qui lissait des comptages (avec une estimation par noyau) et l'affichait avec un gradient d'opacité/transparence. J'ai tout d'abord du passer beaucoup de temps pour comprendre le code d'Etienne afin de l'adapter par la suite. Evidemment, il a fallu rajouter des fonctionnalités, telles la possibilité de modifier l'échelle, changer de ville, modifier le rendu et autres.

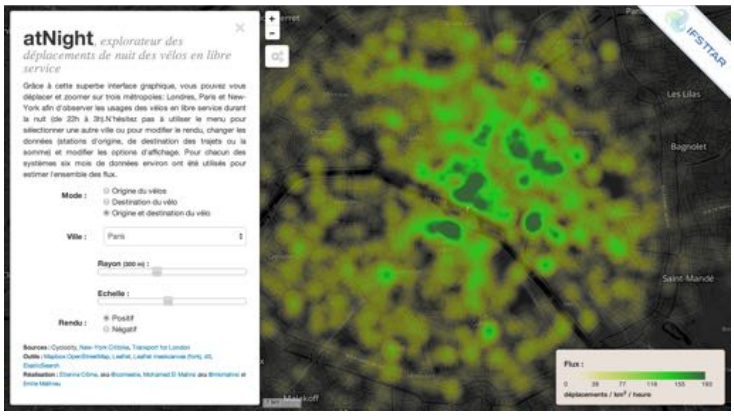

FIGURE 49 – Capture d'écran avec la visualisation atNight

Le résultat est disponible est l'adresse suivante : `mobilletic.ifsttar.fr/mathieue/`

`atNight` (l'expérience est maximale avec la navigateur chrome) ou restylisé à l'adresse suivante : `www.comeetie.fr/galerie/byNight`.

5.2 Projet 2 : TributeToTobler

Le deuxième projet avait pour objectif de visualiser les déplacements des VLS à l'échelle d'une ville et de comprendre en un coup d'oeil quelles sont les mouvements globaux. Nous avons ainsi voulu imiter le site `http://windhistory.com` et sa manière de représenter les déplacements. En effet, ce site permet de visualiser les vents dominants principalement grâce aux données de stations météo des aéroports américains, et ce en affichant de manière circulaire la distribution de la fréquence du vent selon les directions. L'idée a donc été de reprendre cette visualisation partagée sur la page `tributary.io/tributary/3614406`, en affichant la distribution de l'origine ou de la destination des déplacements VLS selon les directions.

Une première étape non triviale de mise en forme des données sur R a été nécessaire. En effet, il fallait agréger les comptages origines et destinations pour chaque station selon les directions. Ainsi, connaissant les coordonnées géographiques, une matrice avec les angles que forment les vecteurs positions des stations (l'origine étant ramenée à chaque station) est générée. Ensuite, la matrice origine destination a été calculée, et ce pour chaque tranche horaire afin de pouvoir visualiser les variations temporelles. Enfin, les comptages ont été agrégés selon les directions (avec une largeur de 10°) puis exportés en format JSON afin d'être lus avec javasript. Après avoir lu, compris puis adapter les fonctions javascript utilisées pour le site windhistory, des fonctions annexes pour changer de villes ou la tranche horaire visualisée ont été implémentées.

Le nom de ce projet : "TributeToTobler" est une référence à l'article de Waldo Tobler (1975) qui propose de représenter les flux par un champs de vecteurs calculé à partir d'une matrice origine destination. Dans ce papier, le modèle fut principalement appliqué à des flux de déplacements humains.

Le résultat est disponible est l'adresse suivante : `mobilletic.ifsttar.fr/mathieue/TributeToTobler`.

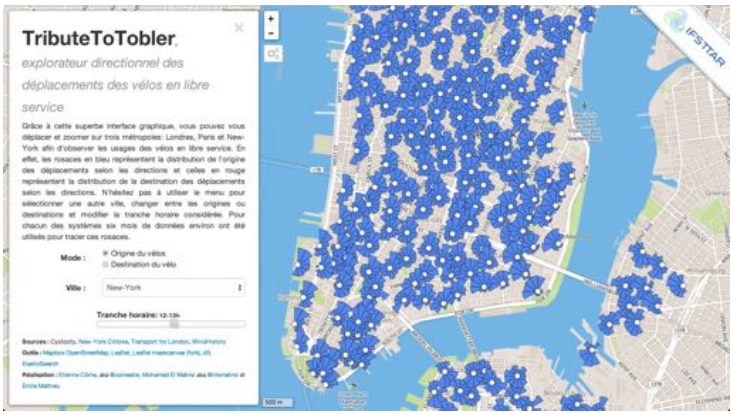

FIGURE 50 – Capture d'écran avec la visualisation TributeToTobler

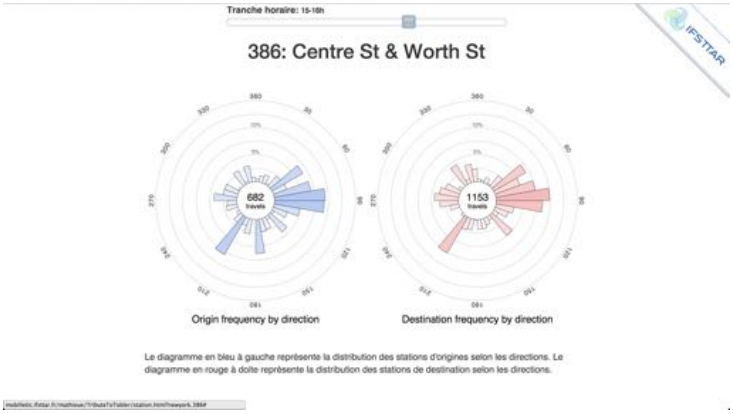

FIGURE 51 – Capture d'écran avec la visualisation TributeToTobler

J'ai été très stimulé par ces mini projets, l'idée de mettre en avant de manière intelligible des données VLS via un site web m'a plu. Mettre en oeuvre un tel site m'a dérouté au début, mais l'aide de mes encadrants ajoutée à de la conviction a finalement abouti à des résultats à peu près convaincants. Beaucoup d'auto critique (et celle de mes encadrants) a aussi été nécessaire afin d'améliorer petit à petit les visualisations.

6 Retour d'expérience

Cette immersion de trois mois dans l'institut de recherche publique qu'est l'Ifsttar fût une expérience fort enrichissante et ce sous plusieurs aspects. J'ai tout d'abord beaucoup appris sur les modes de fonctionnement de la recherche publique, même si tout n'est pas encore clair : financements, hiérarchie, publications,... En outre, le travail collaboratif m'a forcé à m'exprimer en groupe, à lire et comprendre les travaux de collègues et a donc globalement amélioré ma capacité à travailler en équipe. Qui plus est, j'ai assimilé tout au long de ce stage des connaissances sur le machine learning (Apprentissage automatique), l'utilisation du système d'exploitation linux et de langages de programmation tels R et javascript.

6.1 Recherche publique

Le monde de la recherche publique est très spécifique et celui-ci m'étant auparavant inconnu, j'ai pu en apprendre beaucoup.

L'Ifsttar étant un établissement public à caractère scientifique et technologique, il est financé par des capitaux publics, lui permettant entre autre de rémunérer ses chercheurs (fonctionnaires de l'Etat). En outre, des subventions peuvent être attribuées, telle celle du Programme de Recherche et D'Innovation dans les Transports terrestres (PREDIT) qui finance le projet Mobilletic auquel j'ai participé et permet de rémunérer un post-doctorant.

Un aspect très important dans la recherche et dans la vie des chercheurs est la publication d'articles. En effet, celle-ci permet aux chercheurs de communiquer entre eux sur leurs travaux et est aussi un outil d'évaluation. Les publications scientifiques peuvent être distinguées selon leur support de parution : revues scientifiques, comptes-rendus de congrès scientifiques, ouvrages collectifs d'articles de recherche autour d'un thème donné mais mais elles subissent toutes un examen de la rigueur de la méthode scientifique par un comité de lecture indépendant. J'ai également été confronté à la rigueur exigée par la rédaction d'un rapport scientifique et des règles qui le codifient.

J'ai été tout d'abord dérouté par l'aspect de recherche prospective du projet Mobilletic qui ne répond pas à un cahier des charges ou à une question précise posée par une entreprise privée, mais propose une analyse de l'intermodalité par les données mobilité billétique. Il faut donc définir soi-même la ou les problématiques auxquelles se projet permettra de répondre.

6.2 Travail collaboratif

Le travail collaboratif est une composante majeure du travail que j'ai effectué. J'ai ainsi pu apprendre à partager des articles pour construire une bibliographie commune, discuter de la méthodologie à mettre en place et des améliorations souhaitables, à analyser les résultats,...

6.3 Programmation

Durant ce stage j'ai passé la majorité de mon temps devant une machine à coder dans de multiples langages de programmations. J'ai appris à utiliser le langage de statistiques R ainsi que son environnement Rstudio. J'ai aussi pu me familiariser avec Javascript et certaines de ses bibliothèques dans le cadre des projets de visualisations web. J'ai appris à rédiger des documents en LaTeX, ce qui me sera sans doute très utile pour la suite de mon parcours.

7 Conclusion

Nous avons dans ce rapport mis en lumière la manière dont les usagers utilisent le système de vélos en libre-service de Londres. En effet, même si la très grande majorité des usagers de ce système ne l'utilise en réalité que très rarement, les méthodes de classification ont pu faire ressortir des usages typiques.

Cette étude pourrait être complétée par une classification spatiale des stations afin de voir quels groupes de stations sont utilisés par les mêmes usagers. En outre, recouper les groupes d'usagers obtenus par les algorithmes de classification avec des données socio-économiques pourrait permettre de caractériser plus finement ces individus et leurs usages.

Références

Bruno Agard, Catherine Morency, and Martin Trépanier. Mining public transport user behaviour from smart card data. In *12th IFAC Symposium on Information Control Problems in Manufacturing (INCOM)*, 5 2006.

Bruno Agard, Catherine Morency, and Martin Trépanier. Mining smart card data from an urban transit network. In John Wang, editor, *Encyclopedia of Data Warehousing and Mining*, pages 1292–1302. IGI Global, 2009.

David Arthur and Sergei Vassilvitskii. K-means++ : The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 978-0-898716-24-5. URL `http://dl.acm.org/citation.cfm?id=1283383.1283494`.

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.

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

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise Reduction in Speech Processing*, volume 2 of *Springer Topics in Signal Processing*, pages 1–4. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-00295-3. doi : 10.1007/978-3-642-00296-0_5. URL `http://dx.doi.org/10.1007/978-3-642-00296-0_5`.

Alberto Bietti. Latent dirichlet allocation. http ://alberto.bietti.me/files/rapport-lda.pdf, Mai 2012.

Paul Bouman, Evelien van der Hurk, Leo Kroon, Ting Li, and Peter Vervest. Detecting activity patterns from smart card data. In *25th Benelux Conference on Artificial Intelligence (BNAIC 2013)*, 2013.

Park ChangUk. Latent dirichlet allocation, lda. http ://parkcu.com/blog/latent-dirichlet-allocation/, July 2013.

M. Jordan D. Blei, A. Ng. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3 :993–1022, January 2003.

Yves-Alexandre de Montjoye, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the crowd : The privacy bounds of human mobility. *Scientific Reports*, 3 :–, 2013. doi : 10.1038/srep01376.

Takashi 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.

Bertrand Michel Jean-Patrick Baudry, Cathy Maugis. Slope heuristics : Overview and implementation. *Cahier de Recherche Inria*, 7223 :30, Mars 2010.

I.T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

Qi Ju, Alessandro Moschitti, and Richard Johansson. Learning to rank from structures in hierarchical text classification. In *Advances in Information Retrieval*, volume 7814 of *Lecture Notes in Computer Science*, pages 183–194. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-36972-8. doi : 10.1007/978-3-642-36973-5_16. URL `http://dx.doi.org/10.1007/978-3-642-36973-5_16`.

Neal Lathia and Licia 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.

Xiao-lei Ma, Yin-hai Wang, Feng Chen, and Jian-feng Liu. Transit smart card data mining for passenger origin information extraction. *Journal of Zhejiang University SCIENCE C*, 13 (10) :750–760, 2012. ISSN 1869-1951.

Xiao-lei Ma, Yao-Jan Wu, Yin-hai Wang, Feng Chen, and Jian-feng Liu. Mining smart card data for transit riders' travel patterns. *Transportation Research Part C : Emerging Technologies*, 36(0) :1 – 12, 2013.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

Catherine Morency, Martin Trépanier, and Bruno 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.

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Mach. Learn.*, 39(2-3) : 103–134, May 2000. ISSN 0885-6125. doi : 10.1023/A:1007692713085. URL `http://dx.doi.org/10.1023/A:1007692713085`.

Jin Young Park, Dong-Jun Kim, and Yongtaek Lim. Use of Smart Card Data to Define Public Transit Use in Seoul, South Korea, 2008.

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

Quentin Pleplé. Perplexity to evaluate topic models. http ://qpleple.com/perplexity-to-evaluate-topic-models/, May 2013.

Catherine Seaborn, John Attanucci, and Nigel Wilson. Analyzing multimodal public transport journeys in london with smart card fare payment data. *Transportation Research Record : Journal of the Transportation Research Board*, 2121(-1) :55–62, 12 2009.

Waldo Tobler. Spatial interaction patterns. *Journal of Environmental Systems*, 6, 4 :271–301, July 1975.

William 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.