Factorisation de réseaux temporels : étude des rythmes hebdomadaires du système Vélo'v

Ronan HAMON[1], Pierre BORGNAT[1], Cédric FÉVOTTE[2], Patrick FLANDRIN[1], Céline ROBARDET[3]

[1]Laboratoire de Physique (ENS de Lyon, CNRS) Lyon, France

[2]Laboratoire Lagrange (CNRS, OCA & Université Nice Sophia Antipolis) Parc Valrose, Nice, France

[3]LIRIS (INSA de Lyon) Villeurbanne, France

ronan.hamon@ens-lyon.fr, pierre.borgnat@ens-lyon.fr, cfevotte@unice.fr,
patrick.flandrin@ens-lyon.fr, celine.robardet@insa-lyon.fr

Résumé – Nous étudions le système Vélo'v, un système automatisé de location de vélos à Lyon, en le représentant sous la forme d'un réseau temporel. Une décomposition de ce réseau est proposée en utilisant une factorisation en matrices non-négatives (NMF), dont le choix des paramètres est discuté. Cette décomposition permet de représenter à chaque instant le réseau comme un mélange de sous-réseaux, dont les structures décrivent des comportements spécifiques des utilisateurs Vélo'v, en lien avec l'espace géographique et socio-économique. Les coefficients d'activation associés à chacun des motifs permettent de séparer temporellement ceux-ci, permettant une interprétation cohérente avec la littérature sur les vélos en libre-service.

Abstract – We study the Vélo'v system, a fully automated bike-sharing system in Lyon, using a temporal network representation. A decomposition of this network is proposed by using nonnegative matrix factorisation (NMF), whose the choice of parameters is discussed. This decomposition enables us to represent the temporal network as a mixture of subnetwork, whose structures describe specific behaviors of Vélo'v users, in relation with the geographical and socio-economical space. Coefficients of activation associated to the patterns separate temporally each subnetwork, enabling an interpretation consistent with the literature on bike-sharing systems.

Le système Vélo'v : un réseau temporel

FIGURE 1 – Carte de Lyon avec les stations Vélo'v (points bleus) et le diagramme de Voronoï correspondant. En 2011, le système comptait 343 stations actives.

Introduction Le système Vélo'v [1] est un système de vélos en libre-service (VLS) mis en place dans l'agglomération lyonnaise en mai 2005. Le principe repose sur la mise à disposition de vélos dans des stations réparties dans la ville (343 stations actives en 2011, voir Figure 1), où l'utilisateur a la possibilité soit de retirer un vélo, soit d'en déposer un, de manière entièrement automatisée. Les systèmes VLS, et en particulier celui de Lyon, ont déjà fait l'objet de plusieurs études, tant sur les aspects techniques (optimisation du système) que sur les enjeux socio-économiques de ces systèmes, grâce notamment à la mise à disposition de données exhaustives sur les déplacements effectués en 2011. Des approches temporelles [1, 2] ont déjà permis d'identifier des rythmes urbains au cours de la journée et de la semaine, en regardant par exemple le nombre de locations de vélos moyenné sur la semaine (voir Figure 2). Ce travail aborde le système Vélo'v comme un réseau temporel, en proposant une méthode pour extraire automatiquement les caractéristiques spatio-temporelles pertinentes.

Construction du réseau temporel Le réseau temporel est construit à partir des trajets individuels des utilisateurs sur l'année 2011, moyenné sur une semaine. Les nœuds du réseau sont les 343 stations Vélo'v ; un arc (lien dirigé) d'un nœud m vers un nœud n, noté (m, n), correspond à un trajet en Vélo'v de la station m à la station n. Le poids associé à l'arc (m, n) au temps t est défini comme le nombre moyen sur la semaine de trajets pour l'année 2011 partant de la station m vers la sta-

1. http ://www.velov.grandlyon.com/

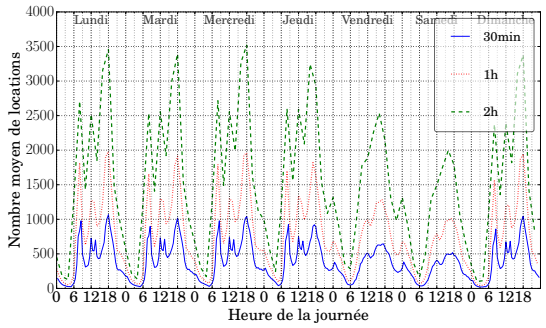


FIGURE 2 – Nombre de locations moyenné sur la semaine pour l'année 2011, pour différents intervalles de temps. L'agrégation des mouvements sur 1h donne un bon compromis entre résolution des détails et fluctuations.

tion n pendant l'intervalle de temps $[t, t + \Delta_t]$, où Δ_t est la période d'agrégation. Le choix de Δ_t a une influence sur le réseau temporel : comme le montre la Figure 2, si Δ_t est petit, la résolution des détails augmente mais les fluctuations sont importantes. Au contraire, choisir un Δ_t élevé réduit les fluctuations au détriment de la résolution. Par la suite, le choix de $\Delta_t = 1$h est retenu, offrant un bon compromis entre fluctuations et résolution. Ce choix amène à considérer 168 intervalles sur une semaine.

Le réseau temporel peut être représenté par un tenseur d'adjacence $\boldsymbol{\mathcal{A}}$, où chaque tranche $\boldsymbol{\mathcal{A}}_{..t}$ du tenseur à l'instant t représente la matrice d'adjacence au temps t. Le tenseur d'adjacence $\boldsymbol{\mathcal{A}}$ a ainsi les dimensions suivantes : $343 \times 343 \times 168$.

Factorisation de la matrice d'adjacence temporelle

La factorisation en matrices non-négatives (*nonnegative matrix factorization* ou NMF) est une technique permettant de décomposer une matrice de données $\boldsymbol{V}$ de dimension $F \times N$ à valeurs non-négatives, en un produit de deux matrices également non-négatives $\boldsymbol{W}$ et $\boldsymbol{H}$ de dimension respectives $F \times K$ et $K \times N$. La NMF permet une réduction de la dimensionnalité des données, en extrayant de façon non supervisée K composantes caractéristiques.

La contrainte de non-négativité des valeurs de $\boldsymbol{W}$ et $\boldsymbol{H}$ induit une représentation "par parties" qui permet d'extraire dans les colonnes de $\boldsymbol{W}$ des motifs caractéristiques des données et dans les lignes de la matrice $\boldsymbol{H}$ les coefficients d'activation de chacun des motifs pour chaque intervalle de temps. Une approche usuelle pour résoudre ce problème consiste en la résolution d'un problème d'optimisation :

$$(\boldsymbol{W}^*, \boldsymbol{H}^*) = \arg\min_{\boldsymbol{WH}} D(\boldsymbol{V}|\boldsymbol{WH})$$

où D est une mesure de dissemblance. S'il n'existe pas de solution analytique à ce problème, un algorithme d'optimisation alterné a été proposé [3] pour le cas où D est la β-divergence, dont les cas particuliers $\beta = 2$ et $\beta = 1$ correspondent respectivement à la distance Euclidienne et à la divergence de Kullback-Leibler.

Plusieurs approches ont été proposées pour adapter la NMF au domaine des réseaux, qu'ils soient statiques [6] ou temporels [4]. Dans cette dernière approche, le tenseur d'adjacence est décomposé par un produit tensoriel de vecteurs de rang 1 à l'aide d'une variante de la NMF appelée factorisation tensorielle non-négatives (NTF). Nous proposons ici de transformer le tenseur d'adjacence en une matrice et d'appliquer la NMF classique sur cette matrice. Pour cela, pour chaque intervalle de temps $[t, t + \Delta_t]$, la matrice d'adjacence correspondante est transformée en un vecteur colonne par empilement bout-à-bout des colonnes. Ce vecteur constitue la $t$ème colonne de la matrice $\boldsymbol{V}$. Après décomposition par NMF, les colonnes de $\boldsymbol{W}$ sont ainsi des matrices d'adjacence, obtenues suivant le procédé inverse.

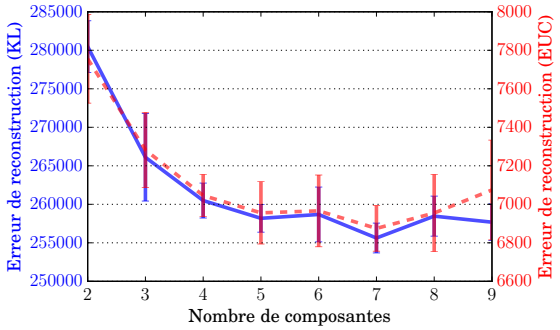


FIGURE 3 – Erreur de reconstruction après validation croisée pour K variant de 2 à 9. L'erreur est calculée en utilisant la divergence de Kullbach-Leibler (solide bleue) et la distance Euclidienne (pointillés rouges).

Deux paramètres nécessitent d'être réglés dans l'algorithme de la NMF : la mesure de dissemblance D, contrôlée ici par la valeur de β, et le nombre de composantes K. Le choix de β est guidé par le modèle probabiliste dans lequel nous nous plaçons par rapport à la nature des données : dans notre application sur le système Vélo'v, les éléments v_{it} de la matrice $\boldsymbol{V}$ décrivent un nombre moyen de vélos au départ d'une station vers une autre station pendant un intervalle de temps, qui peuvent se modéliser par une loi de Poisson : $v_{it} \sim \mathcal{P}(v_{it}, \sum_{k=1}^{K} w_{ik} h_{kt})$, où $\mathcal{P}(x, \lambda) = e^{-\lambda} \frac{\lambda^x}{\Gamma(x+1)}$ avec $\Gamma(x+1)$ la fonction Gamma. Comme décrit dans [5], sous des hypothèses d'indépendance des éléments de $\boldsymbol{W}$ et $\boldsymbol{H}$, maximiser la vraisemblance du modèle revient à minimiser la divergence de Kullback-Leibler, c'est-à-dire la β-divergence pour $\beta = 1$. Il n'existe en revanche pas de choix naturel pour le nombre de composantes K, que nous proposons de choisir par validation croisée : pour une valeur de K fixée, la NMF est réalisée sur une matrice $\boldsymbol{V}_{VC}$ construite en supprimant aléatoirement 50% des valeurs de $\boldsymbol{V}$. La matrice $\tilde{\boldsymbol{V}}$ obtenue est comparée avec les valeurs de $\boldsymbol{V}$ sur les données manquantes de $\boldsymbol{V}_{VC}$, en utilisant la distance Euclidienne et la divergence de Kullback-Leibler. Ce processus est répété 15 fois afin d'évaluer la dispersion de l'erreur de recons-

truction. La Figure 3 affiche l'erreur de reconstruction pour K variant de 2 à 9, afin de conserver un nombre raisonnable de motifs dans l'interprétation. Pour les deux fonctions d'erreur, la valeur de K qui minimise l'erreur de reconstruction est $K = 7$, choix qui sera retenu par la suite pour la décomposition du réseau temporel.

Analyse temporelle des rythmes hebdomadaires

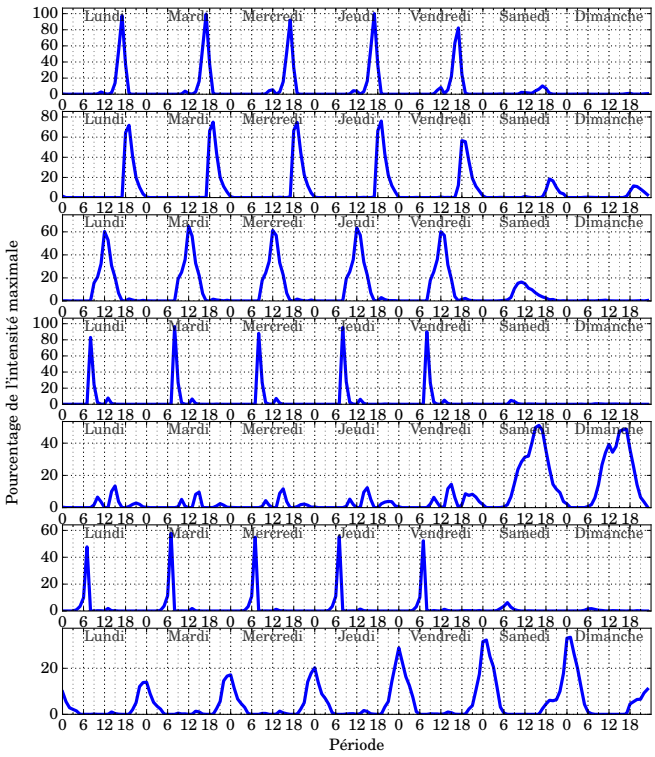


FIGURE 4 – Coefficients d'activation pour chaque motif au cours de la semaine.

La Figure 4 affiche pour chacun des motifs les coefficients d'activation pour chaque intervalle au cours de la semaine. Les pics d'intensité permettent de classer les composantes suivant deux critères : la période de la journée (matin, midi, après-midi, soirée et nuit) et le type de jour (jour de semaine ou week-end), ce qui est réalisé dans la Table 1. La description temporelle des composantes montre que la NMF extrait de manière automatique les périodes pertinentes d'un point de vue socio-économique, à savoir les pics d'activité du système explicités par la Figure 2.

On peut noter que la modulation du coefficient d'activation permet d'associer un même comportement, décrit dans la section suivante, avec des variations dans l'intensité de ce comportement. Par exemple, la composante 7 exprime une activité présente la nuit entre 23h et 3h du matin, similaire pour chaque jour de la semaine, mais avec une intensité plus forte le jeudi, vendredi et samedi, jours où l'activité nocturne est plus intense.

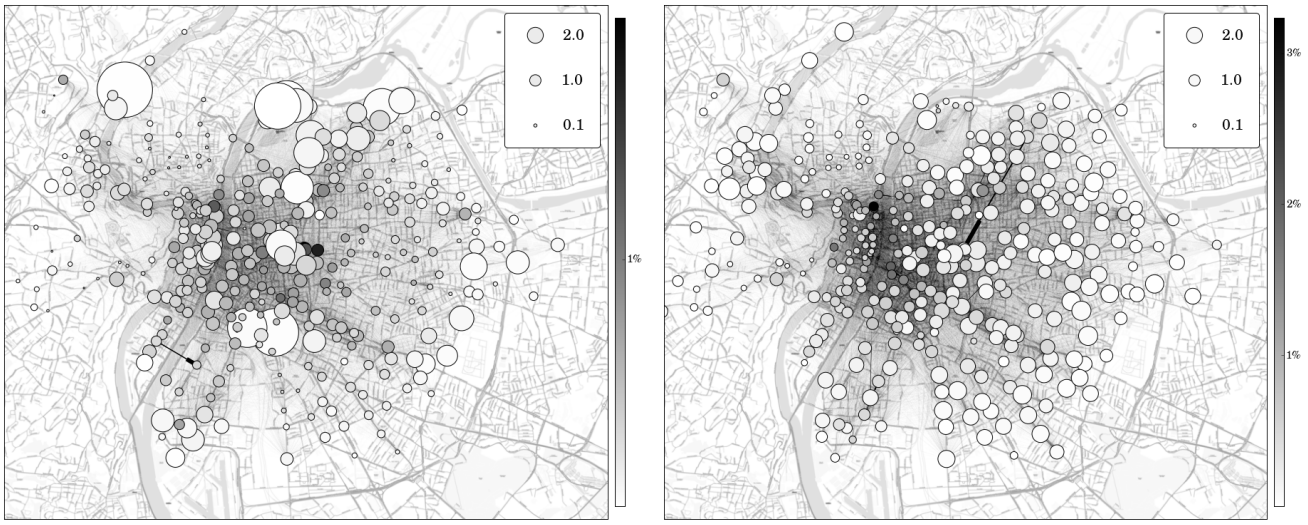
Motif	Période
1	Semaine - Après-midi (17h)
2	Semaine - Début de soirée (18h-19h)
3	Semaine - Midi (11h-14h)
4	Semaine - Matin (8h)
5	Week-end - Journée
6	Semaine - Matin (7h)
7	Nuit (23h-3h)

TABLE 1 – Classification des composantes en fonction des pics d'intensité des coefficients d'activation suivant deux critères : la période de la journée (matin, midi, après-midi, soirée et nuit) et le type de jour (jour de semaine ou week-end).

Analyse spatiale des rythmes hebdomadaires

Les motifs obtenus se présentent sous la forme d'une matrice d'adjacence. Par souci de visualisation, la Figure 5 affiche ces matrices sous la forme d'un graphe déployé dans l'espace géographique lyonnais, où les nœuds du graphe sont les stations. Le degré sortant w_{out} (respectivement le degré entrant w_{in}) d'un nœud est défini comme la somme des poids des liens partant du nœud (respectivement la somme des poids des liens arrivant vers le nœud). La couleur indique le degré sortant du nœud sur une échelle de blanc à noir. Comme les motifs sont normalisés, ce degré représente un pourcentage de l'activité totale. La taille du point indique le ratio $\frac{w_{\text{in}}}{w_{\text{out}}}$ pour chaque nœud : plus le nœud est grand, plus la station se remplit. Inversement, plus le nœud est petit, plus la station se vide.

L'interprétation spatiale des motifs est ici limitée à deux composantes. La Figure 5a affiche le réseau correspondant au motif 4, principalement actif le matin les jours de semaine. L'étude de ce réseau permet de retrouver des éléments déjà connus sur les rythmes matinaux dans la ville de Lyon. Tout d'abord, l'activité est concentrée sur quelques stations dans des zones spécifiques telles que Part-Dieu, regroupant la gare, un centre commercial et le quartier des affaires, ainsi que dans le centre de la presqu'île. Les comportements des stations sont également cohérents avec les analyses socio-économiques déjà réalisées : les stations qui se remplissent sont situées autour des campus (La Doua, Université Jean Moulin, etc.) ou des zones à forte activité commerciale. Parallèlement, les stations qui se vident sont situées dans les zones résidentielles (8ème arrondissement, ouest du 3ème arrondissement, etc.) et les zones en altitude (Croix-Rousse, Fourvière). Ce réseau souligne ainsi la disparité géographique des déplacements urbains dans une période de grande activité. Les mouvements sont mieux répartis pendant les périodes de faible activité, comme le montre la Figure 5b correspondant au motif 7, dont les coefficients d'activation présentent des pics d'intensité la nuit entre 23h et 3h du matin chaque jour de la semaine. Si l'activité est concen-



(a) Motif 4

(b) Motif 7

FIGURE 5 – Représentation des motifs sous la forme d'un graphe déployé dans l'espace géographique lyonnais. Les nœuds correpondent aux stations. La couleur des nœuds indique le degré sortant sur une échelle de blanc à noir, représentant un pourcentage de l'activité totale. La taille des nœuds indique le ratio entre degré entrant et degré sortant. Les arcs entre les nœuds sont représentés et leur épaisseur est proportionnelle au poids associé à chaque arc.

trée comme attendu autour des zones dynamiques telles que les quartiers de l'Opéra, de Saint Jean ou de la Guillotière, regroupant de nombreux bars et restaurants, les stations situées au centre sont enclines à se vider au profit de celles situées à la périphérie de la ville. Cela souligne ainsi l'usage du vélo comme moyen de transport alternatif lorsque les réseaux de métro et bus ferment.

Conclusion

Les résultats de ce travail ont ouvert la voie à une analyse conjointe de la dynamique d'un réseau temporel et des motifs d'interactions caractérisant cette dynamique. Un des objectifs est de pouvoir séparer en temps et dans le domaine des graphes les caractéristiques du réseau temporel. L'application de cette méthode sur des données réelles issues du système Vélo'v, dont la représentation sous la forme d'un réseau est naturelle, a permis de mettre en évidence des comportements connus, à la fois des analyses quantitatives et des études socio-économiques déjà réalisées, de manière totalement non supervisée. Cette étape valide notre approche, et nous encourage à étendre l'analyse des résultats pour rechercher les comportements inattendus, afin de pouvoir tester sociologiquement des hypothèses plus subtiles et moins facilement observables.

Remerciements Ce travail a été réalisé avec le soutien des programmes ARC 5 et ARC 6 de la région Rhône-Alpes, du GdR ISIS et du projet Vél'Innov ANR-12-SOIN-0001-02.

Références

[1] Pierre Borgnat, Patrice Abry, Patrick Flandrin, Céline Robardet, Jean-Baptiste Rouquier, and Eric Fleury. Shared bicycles in a city : a signal processing and data analysis perspective. *Advances in Complex Systems*, 14(03) :415–438, 2011.

[2] Pierre Borgnat, Céline Robardet, Patrice Abry, Patrick Flandrin, Jean-Baptiste Rouquier, and Nicolas Tremblay. A dynamical network view of lyon's vélo'v shared bicycle system. In *Dynamics On and Of Complex Networks, Volume 2*, Modeling and Simulation in Science, Engineering and Technology, pages 267–284. Springer New York, 2013.

[3] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9) :2421–2456, June 2011.

[4] Laetitia Gauvin, André Panisson, and Ciro Cattuto. Detecting the community structure and activity patterns of temporal networks : A non-negative tensor factorization approach. *PLoS ONE*, 9(1) :e86028, 2014.

[5] Tuomas Virtanen, Ali Taylan Cemgil, and Simon Godsill. Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 1825–1828. IEEE, 2008.

[6] Zhongyuan Zhang, Chris Ding, Tao Li, and Xiangsun Zhang. Binary matrix factorization with applications. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 391–400. IEEE, 2007.