



CluCo, EGC 2014

Analyse des trajets de Vélib par clustering

Yousra Chabchoub – Institut Supérieur d'Electronique de Paris
Christine Fricker – INRIA Rocquencourt

Plan

- ▶ Contexte et problématiques liées aux systèmes de vélos en libre service
- ▶ Description de l'algorithme de clustering Kmeans
- ▶ Expérimentations
 - ▶ Présentation des données
 - ▶ Classification en clusters
 - ▶ Identification des stations problématiques
- ▶ Conclusion et perspectives

Histoire des VLS (Vélos en Libre Service)

- ▶ 1965 - Amsterdam: vélos peints en blanc, complètement libres
- ▶ 1974 - La Rochelle propose 350 vélos jaunes sur 3 points de location
- ▶ 1991 - Copenhague
- ▶ 1998 - Rennes
- ▶ 2005 - Lyon
- ▶ ... un peu partout dans le monde (Europe, Amérique, Asie, Australie), plus de 500 000 vélos
- ▶ La Chine (Hangzhou): le plus grand système de VLS: 50 000 vélos et 2 000 stations (une tous les 100 m)

Vélib en quelques chiffres

- ▶ Début: 15 juillet 2007
- ▶ Service disponible 24 heures sur 24 et 7 jours sur 7
- ▶ 20 000 vélos
- ▶ 1 230 stations, une station tous les 300 mètres environ
- ▶ 110 000 trajets par jour en moyenne
- ▶ 3 000 vélos déplacés par les camions par jour
- ▶ Tarifs
 - ▶ Les abonnements annuels
 - Vélib' Classique - 29 € → 30 premières minutes offertes
 - Vélib' Passion - 39 € → 45 premières minutes offertes
 - ▶ Les tickets courte durée, 30 premières minutes offertes à chaque déplacement
 - 1 jour - 1,70€ , 7 jours - 8€

Problématiques liées à Vélib



- ▶ Comment améliorer la disponibilité des ressources ?
(Stations complètement vides et stations saturées)
 - ▶ Méthodes incitatives
 - ▶ Méthodes régulatrices
 - Dégager une typologie des usages dans un but d'amélioration du service aux usagers

État de l'art: travaux de recherche autour de VLS

- ▶ Fricker et Gast 2010: Établissement d'un nombre de vélos optimal par station dans un réseau homogène:
 $n = \text{capacité}/2 + \text{taux de départ}/\text{taux d'arrivée}$
- ▶ Fricker et al. 2012: Extension des résultats au cas d'un réseau inhomogène
- ▶ Thèse de Daniel Chemla 2012
 - ▶ Optimisation de la régulation par les camions
 - ▶ Méthodes incitatives basées sur des tarifs calculés en temps réel
- ▶ Fabio Pinelli 2012
 - ▶ Etablissement d'un modèle de prédiction de l'état des stations

État de l'art: Vélib et clustering

▶ Objectif du clustering

- ▶ Chercher des similarités entre les activités ou les états des stations au cours du temps
- ▶ Dégager une typologie des usages dans un but d'amélioration du service aux usagers.

▶ Exemples de travaux

- ▶ Côme et Oukhellou 2012, IFSTTAR
 - ▶ Un modèle statistique de clustering, données de trajets de Vélib
- ▶ Froehlich et al. 2009, Université de Maryland
 - ▶ Hierarchical clustering, données de trajets de Barcelone
- ▶ Borgnat et al. 2011, ENS Lyon
 - ▶ Identification des groupes de stations qui échangent des vélos, données de trajets de Vélo'v

Les algorithmes de clustering

- ▶ Deux algorithmes de clustering non supervisés largement utilisés
 - ▶ K-means
 - ▶ Hierarchical clustering
- ▶ Autres méthodes de clustering
 - ▶ Fuzzy C-Means
 - ▶ Analyse en Composante Principale (ACP)

Description de Kmeans 1/2

- ▶ Un algorithme simple basé sur un apprentissage non supervisé (James MacQueen, 1967)
- ▶ Parfaitement adapté aux données qui comportent la notion de centre
- ▶ Le nombre de classes K doit être préalablement choisi
- ▶ Principe de Kmeans
 - ▶ Chaque objet est représenté par un vecteur de \mathbb{R}^P et est associé au centre de la classe la plus proche au sens d'une métrique préalablement définie
 - ▶ L'affectation de tous les objets et la mise à jour des centres de classes seront répétées jusqu'à la convergence

Description de Kmeans 2/2

- ▶ La distance euclidienne est la métrique la plus utilisée
→ centres de classes = centres de gravité
- ▶ Initialisation des centres des clusters
 - ▶ Choisir K points avec des coordonnées aléatoires → peut engendrer des clusters vides
 - ▶ Choisir au hasard K points parmi l'ensemble des données à partitionner (Pakhira 2009)

Description des données

- ▶ Données fournies par la ville de Paris et JCDecaux dans le cadre de la politique de l'Open Data.
- ▶ 6 mois de trajets: avril → juillet, octobre → décembre 2013
- ▶ Un trajet = tempsD, tempsA, stationD, stationA, typeTrajet (maintenance, régulation, vrai trajet)
- ▶ 2 grandes catégories: Jours travaillés / Week-end
- ▶ Une forte périodicité
- ▶ Analyse des trajets du vendredi 28/06/2013
 - ▶ 121 709 trajets
 - ▶ 1.03% maintenance
 - ▶ 1.48% régulation

Application de Kmeans aux données

- ▶ Pour chaque station i , on définit sa série temporelle relative à ses trajets:

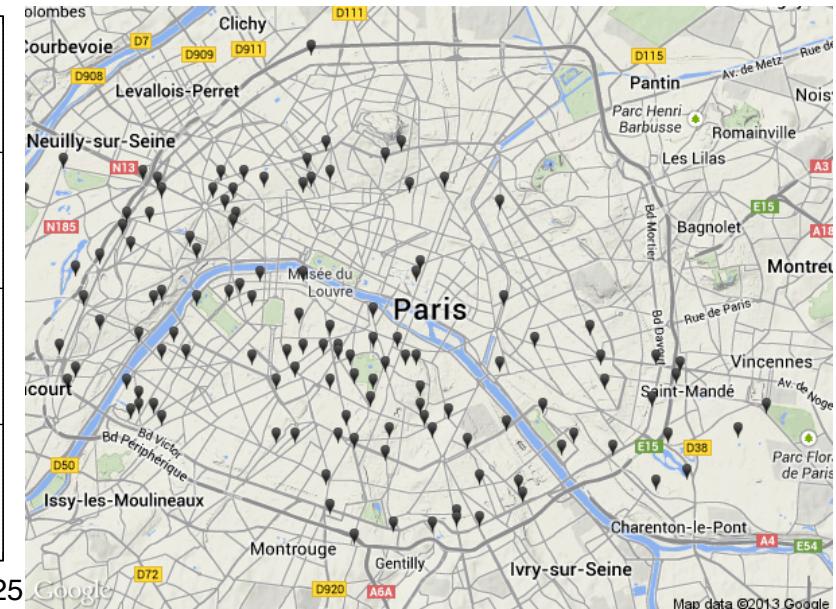
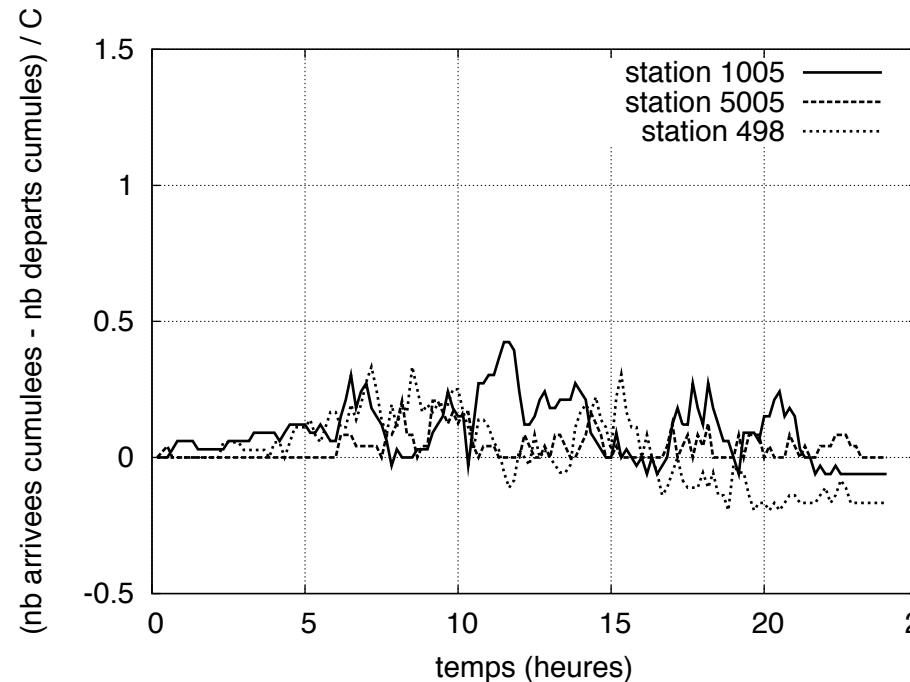
$$T_{i,t} = (a_{i,t} - d_{i,t}) / C_i$$

- ▶ $a_{i,t}$: nombre d'arrivées cumulées à la station i (entre 0h et t)
- ▶ $d_{i,t}$: nombre de départs cumulés depuis la station i
- ▶ C_i : capacité de la station (nombre de bornes, $\in [8, 72]$)
- ▶ $T_{i,t}$ est calculée toutes les 10 minutes
- ▶ En l'absence de régulation, on a $T_{i,t} \in [-1, 1]$
- ▶ Le nombre de clusters $K=6$

Résultats: les clusters

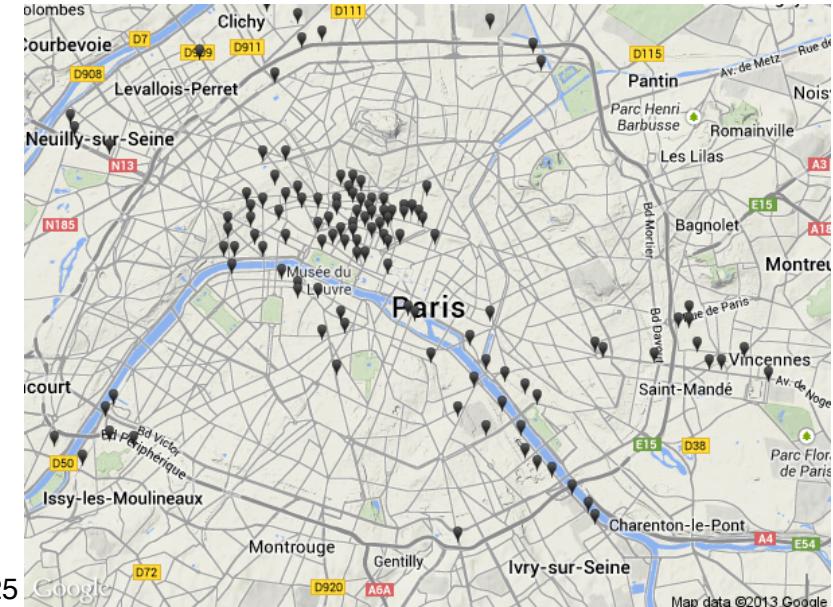
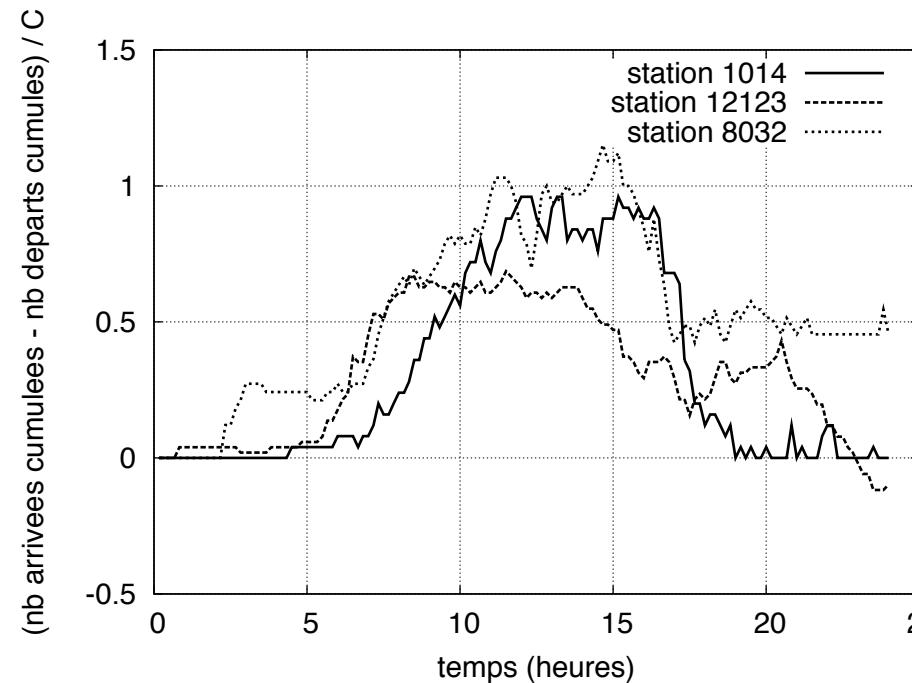
- ▶ Initialisation aléatoire des clusters → résultats moyennés sur 20 exécutions de l'algorithme
 - ▶ Les tailles et contenus des clusters sont très peu variables
 - ▶ Temps moyen d'exécution : 585 ms
 - ▶ Nombre moyen d'itérations : 26.35
- ▶ Les clusters sont très similaires à ceux présentés par Côme et Oukhellou en 2012

Cluster mixte: 500 stations



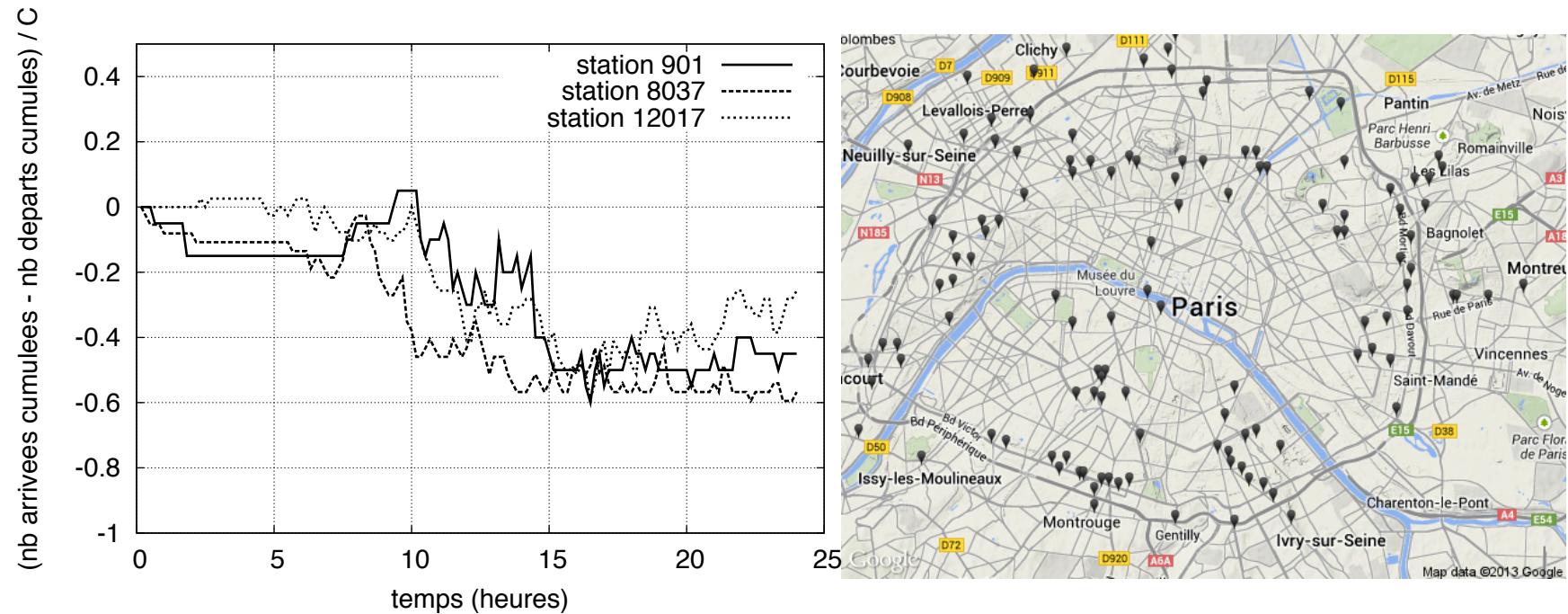
- ▶ Stations équilibrées sur toute la journée

Cluster emploi : 177 stations



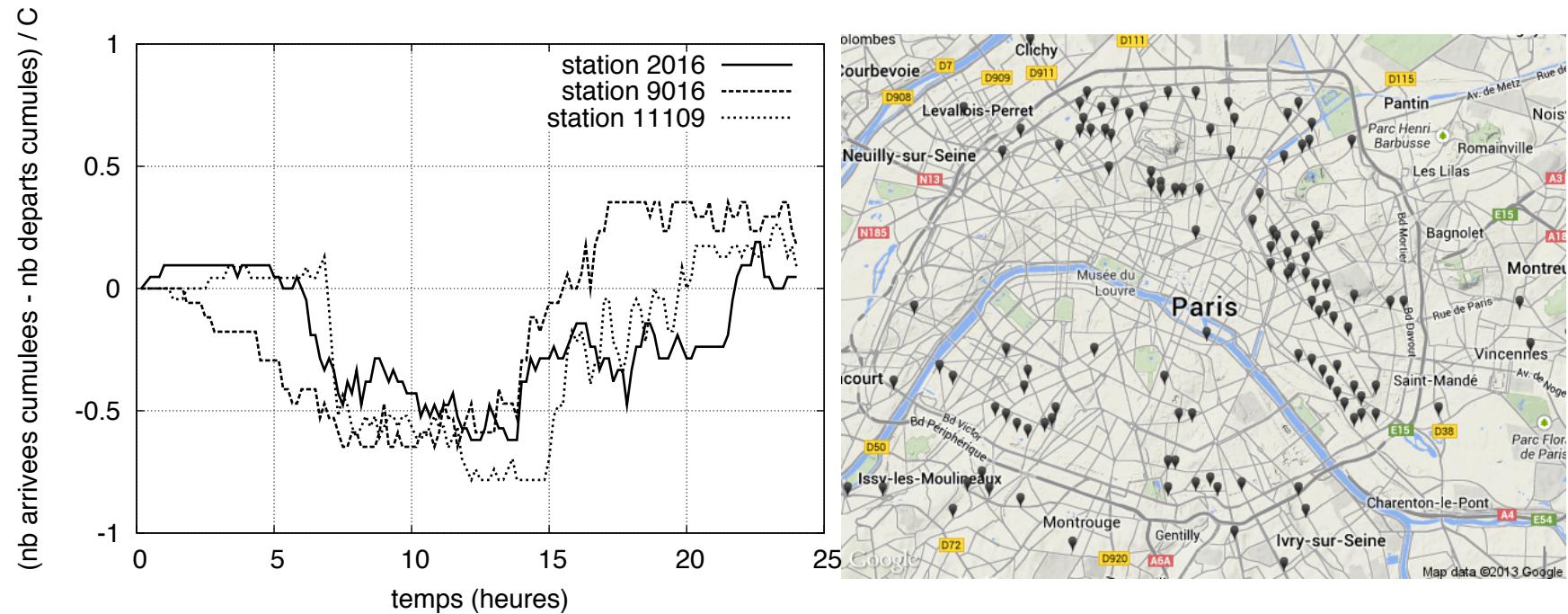
- ▶ Un grand nombre d'arrivées entre 6h et 10h
- ▶ Un grand nombre de départs entre 16h et 18h

Cluster périphérie : 177 stations



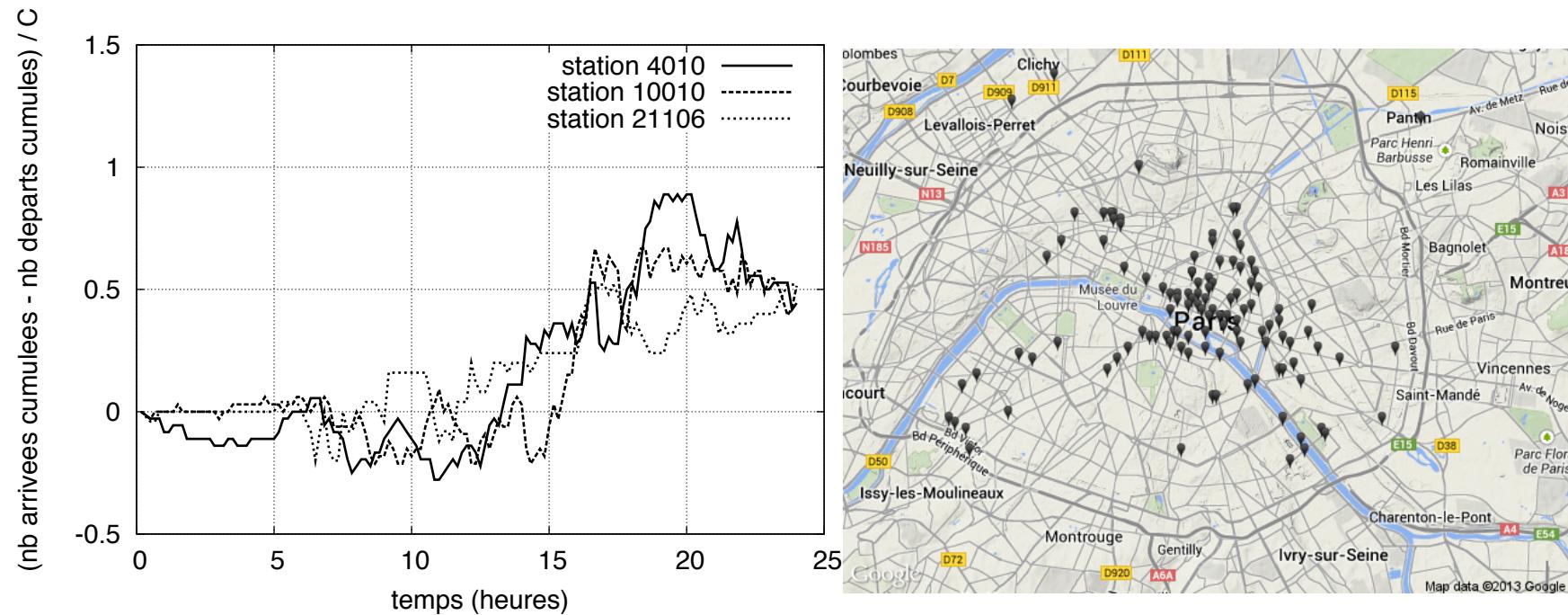
- ▶ Beaucoup de départs entre 10h et 15h

Cluster habitations : 169 stations



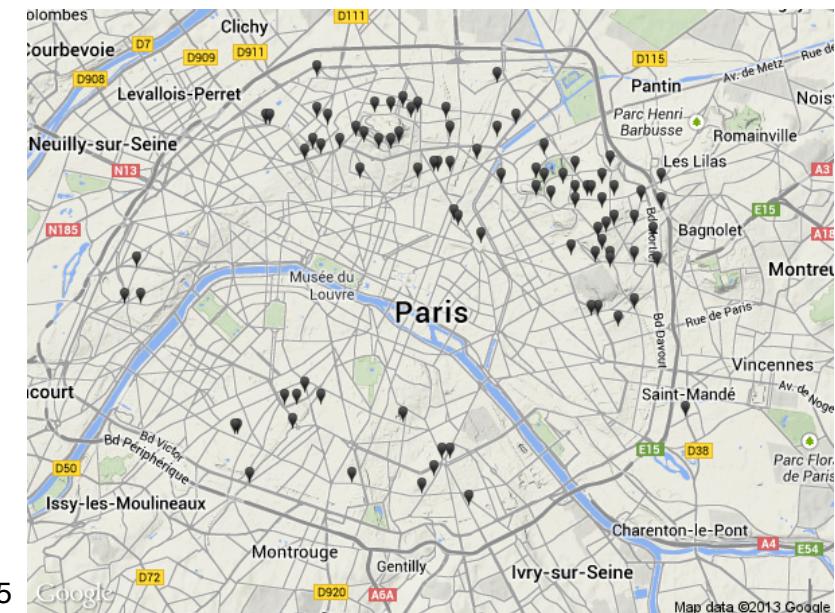
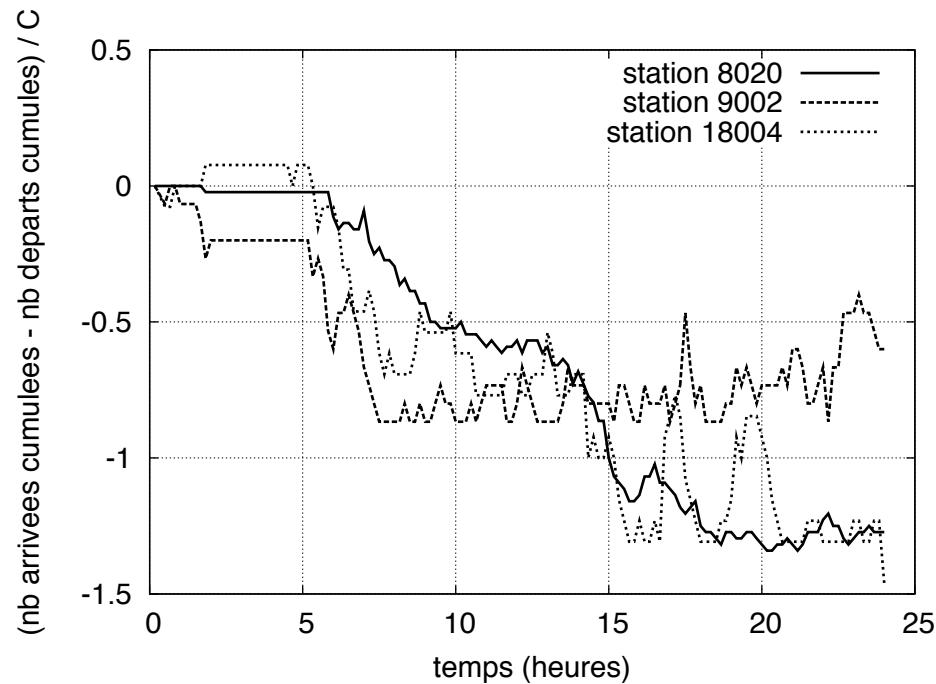
- ▶ Beaucoup de départs entre 6h et 8h
- ▶ Beaucoup d'arrivées entre 15h et 18h

Cluster divertissement : 115 stations



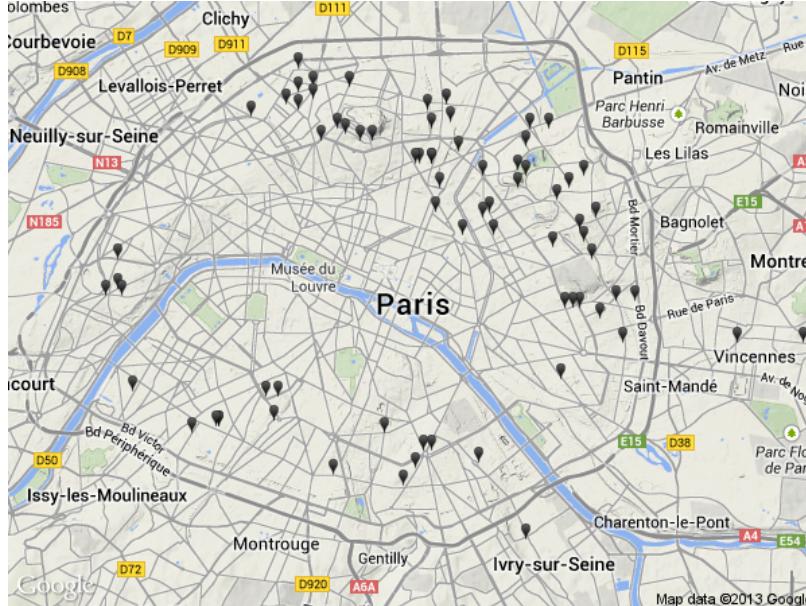
- ▶ Des arrivés massives le soir à partir de 16h

Cluster gares : 87 stations



- ▶ Stations souvent vides

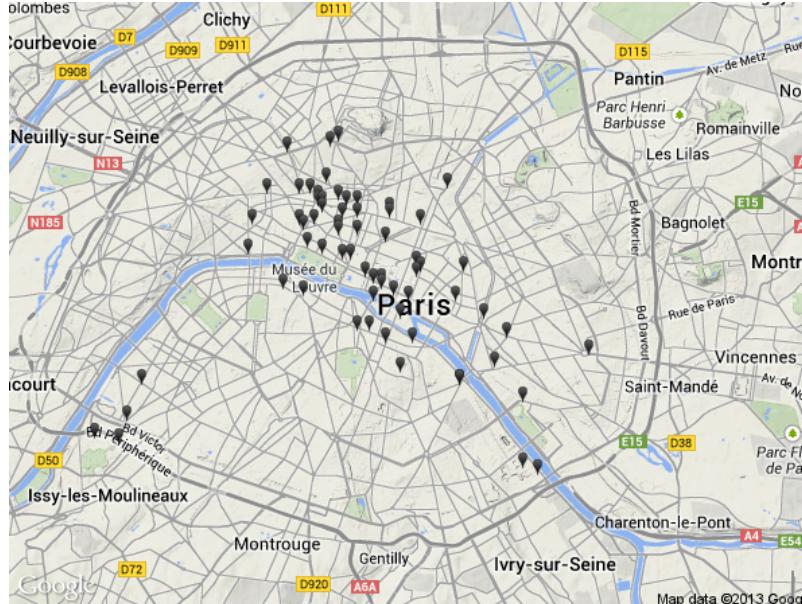
Les stations vides



Stations vides

- ▶ Définition: station vide, $T_{i,t}$ atteint -0.9 à un instant t de la journée
- ▶ 73 stations vides
- ▶ Composition: clusters Habitation (30%) et Gares (65%)

Les stations pleines



Stations pleines

- ▶ Définition: station pleine, $T_{i,t}$ atteint 0.9 à un instant t de la journée
- ▶ 66 stations pleines
- ▶ Composition: clusters Divertissement (50%) et Emploi (40%)

Conclusion et perspectives

- ▶ Etude exploratoire des données des trajets Vélib pour une meilleure compréhension de l'utilisation de ce système
- ▶ Perspectives
 - ▶ Tester d'autres algorithmes de clustering (Hierarchical clustering...) et comparer leurs performances à celles de Kmeans
 - ▶ Etudier l'impact d'une station attractive sur son voisinage
 - ▶ Faire des simulations pour répondre aux questions suivantes
 - ▶ Quel est l'impact d'une augmentation de ressources (nombre de bornes, nombre de vélos) ?
 - ▶ Quel est l'impact de la régulation ?
 - ▶ Par quelles méthodes incitatives les utilisateurs peuvent-ils être amenés à mieux répartir les vélos entre les stations ?

Autres réflexions

- ▶ Comment étendre le système Vélib à la banlieue parisienne ?
- ▶ Faut-il modifier la cartographie de Vélib en période de vacances ?
- ▶ Quel est l'impact des évolutions des autres moyens de transport sur le système Vélib ?