La classification automatique des données billettiques : une application au réseau de transports lyonnais

C. Bayart*, B. Cottreau**, D. Clot*, P. Bonnel**

* LSAF, Université Lyon 1, France
** LAET, Université Lyon 2, France

caroline.bayart@univ-lyon1.fr

19 mai 2022

Table des matières

Contexte et méthodologie

Motivations de la recherche

- De récents bouleversements ont impacté la mobilité individuelle (Borkowski et al., 2021 ; Deschaintres, 2018)
  - Les pouvoirs publics questionnent l'impact de la Covid 19 sur la fréquentation du réseau (Guelton et Poinsot, 2020)
- Les données billettiques sont une opportunité intéressante (Pelletier et al., 2011) :
  - collecte en continu et de manière passive
  - caractérisation temporelle fine des comportements
- Des études sont en cours dans plusieurs métropoles (Bourdeau et Morency, 2022)

Autres données, autres méthodes

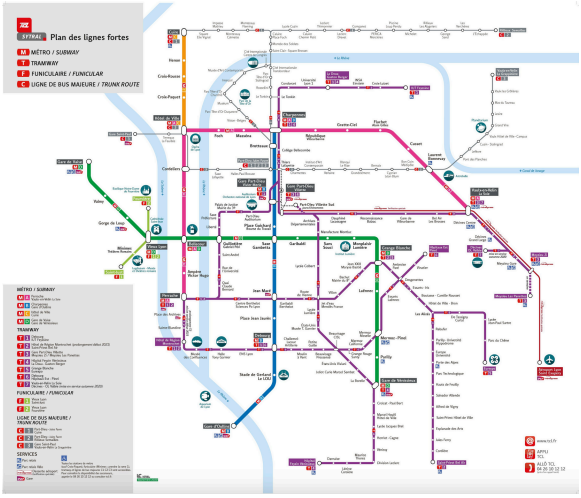
- Les méthodes d'analyse multidimensionnelles sont particulièrement adaptées à ces fichiers complexes
  - ACP : synthétiser les variables et éliminer le bruit
  - clustering : identifier des structures au niveau des individus et les caractériser avec les variables
- Ces outils connaissent un nouvel essor en économie des transport
  - variabilité intra personnelle de l'usage des TC (Egu et Bonnel, 2020)
  - régularité des comportements de mobilité (Goulet-Langlois et al., 2018)

=> analyser l'évolution de la fréquentation des TC (métro, bus et tram) dans une perspective temporelle (2015-21) et spatiale

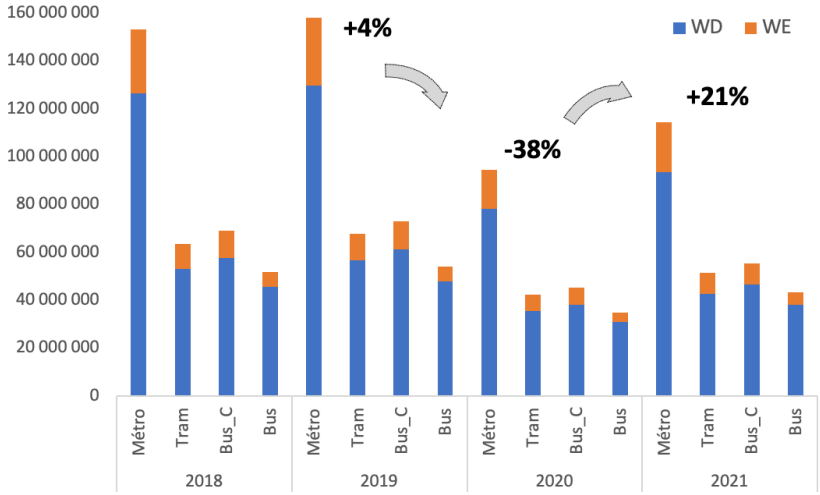
Application au réseau de TC lyonnais

Un réseau dense et diversifié

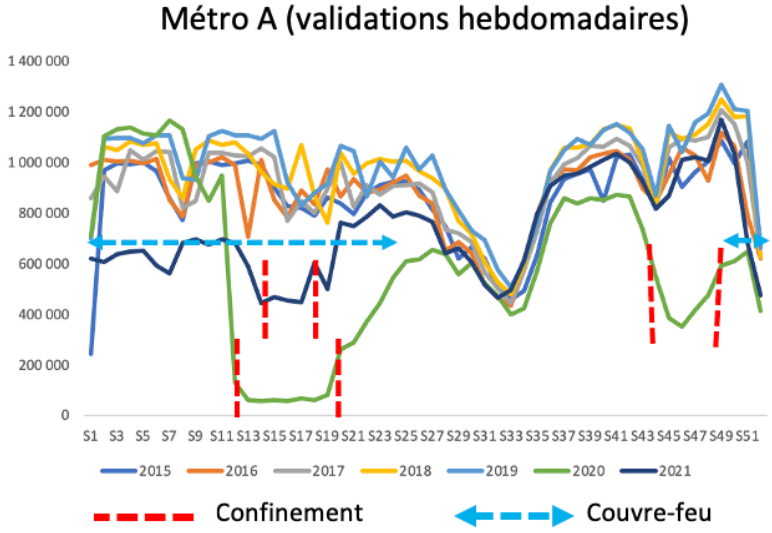
Kéolis exploite 4 métros (+2 funic.), 7 trams et plus de 120 bus (1.8 M voy./j).



Une évolution contrastée de la fréquentation



L'impact de la crise sanitaire



L'originalité du jeu de données

Un vecteur d'intensité hebdomadaire des validations est calculé :

- pour chaque station de métro, ligne de tram et de bus
- pour chaque année étudiée (2015 à 2021)

Intensité hebdomadaire des validations

Pour chaque année, la semaine 2 est choisie comme période de référence.

$$intensite_j = validations_j / validations_2 \quad (1)$$

L'originalité du jeu de données

		Intensité des validations			
Station	Année	S2	S3	...	S51
AMPERE	2015	1	1.06	...	1.20
AMPERE	2016	1	1.00	...	0.77
AMPERE	2017	1	0.88	...	1.03
AMPERE	2018	1	1.04	...	1.23
AMPERE	2019	1	0.99	...	1.21
AMPERE	2020	1	0.99	...	0.58
AMPERE	2021	1	1.01	...	1.16

L'originalité du jeu de données

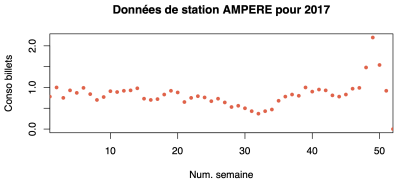
Plusieurs jeux de données sont à analyser :

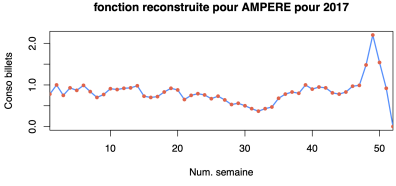
- Selon le mode de transport (bus, métro, tram…)
- Selon un mode d'agrégation à la semaine ou différenciant le week-end des autres jours

Nous présentons les résultats obtenus sur l'analyse des validations du métro en semaine complète ou en semaine hors week-end.

Autres données, autres méthodes

- Les données possèdent une structure sous-jacente (variation temporelle), qui peut être représentée par une ou plusieurs fonctions.
- Les séries annuelles seront considérées comme des discrétisations de fonctions continues du temps, approchées par des fonctions affines par morceaux.
- Les individus étudiés sont vus comme des fonctions (Ramsey et Silverman, 2005), non comme des vecteurs de $\mathbb{R}^n$.

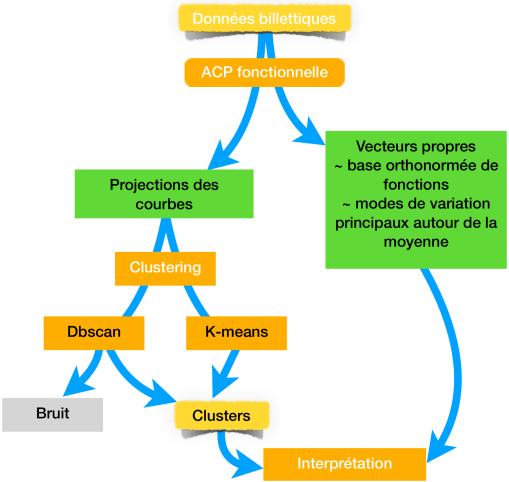




Autres données, autres méthodes

- La base de fonctions orthonormées (vecteurs propres) déterminée par l'ACP fonctionnelle permet d'exprimer les séries comme des points de l'espace vectoriel $L_2^n(T)$. Cette base est utilisée pour interpréter les variations principales autour de la moyenne des courbes projetées dans l'espace factoriel.
- Dans cet espace, nous cherchons à caractériser des groupes avec deux méthodes de clustering (k-means et dbscan).
- Les clusters peuvent être observés dans les plans factoriels, mais également sur les nuages de courbes.

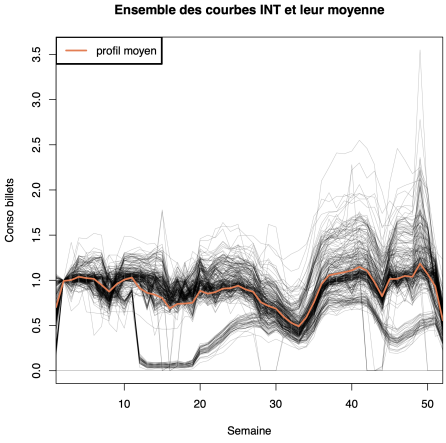
Méthode de construction et d'interprétation des clusters



Résultats et illustrations

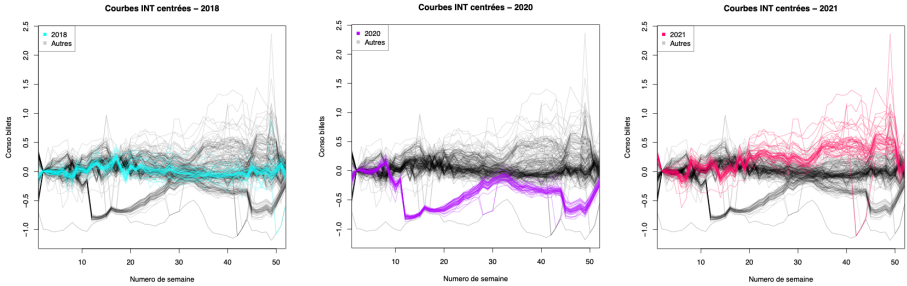
Analyse des profils - Semaines complètes

Aperçu de l'ensemble des courbes d'intensité et de leur moyenne



Analyse des profils - Semaines complètes

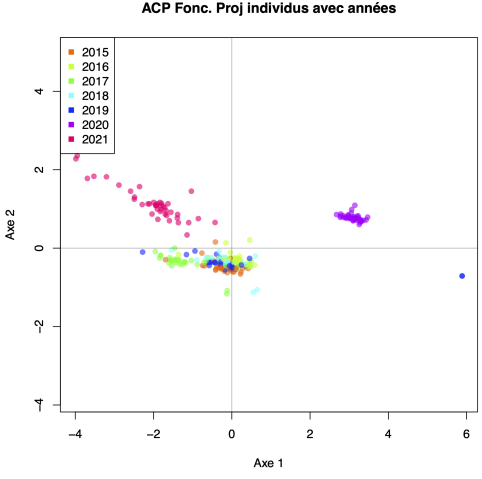
Intensité par station, centrée autour de l'intensité moyenne par semaine (toutes stations et toutes années)



2020 (COVID-19) est caractérisée par des courbes sous le niveau moyen alors que 2021 semble montrer un effet de rattrapage

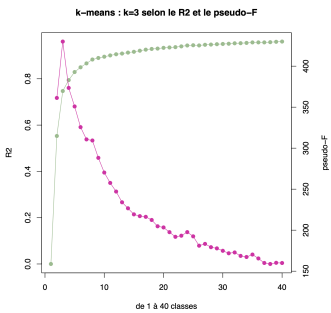
Projection des individus - Semaines complètes

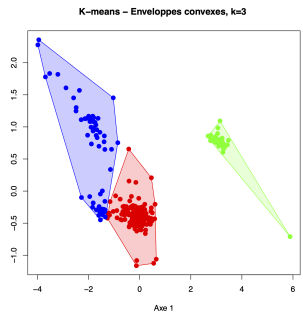
1e plan factoriel (+80% iner.) : nous retrouvons ces singularités de 2020 et 2021

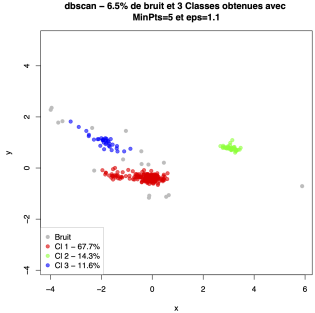


Etude des clusters - Semaines complètes

La recherche du nombre de clusters par les critères classiques conduit à 3.

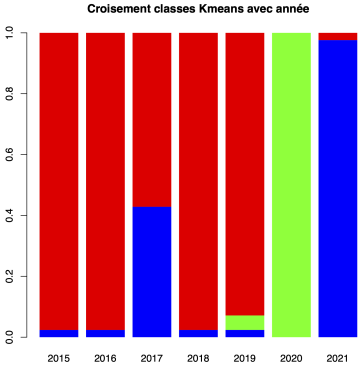


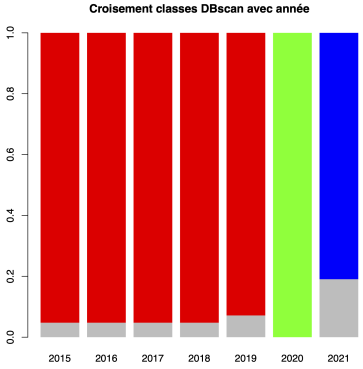




Répartition des clusters - Semaines complètes

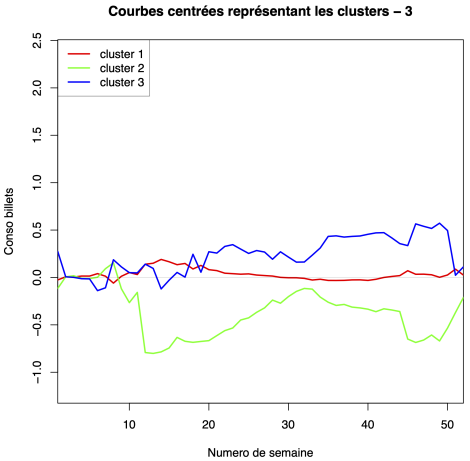
Quelle que soit la méthode retenue, les profils des années 2020 et 2021 sont assez particuliers pour constituer 2 groupes distintcs.



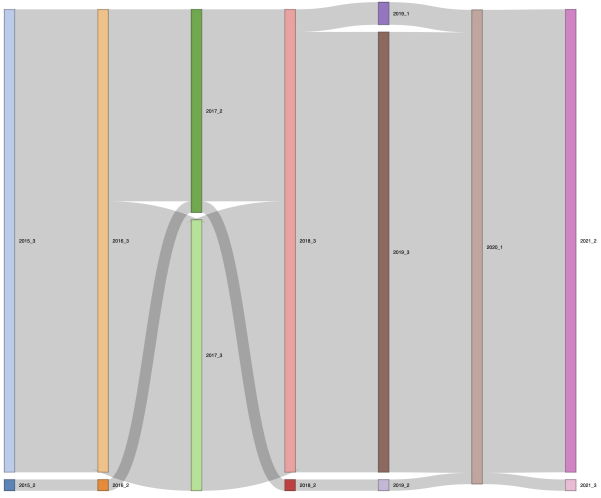


Variation des données d'intensité - Semaines complètes

- cl1 : peu d'écarts à la moyenne
- cl2 (2020) : baisse importante au confinement 1 (printemps), puis reprise progressive jusqu'au confinement 2 (rentrée) et reprise lente en fin d'année
- cl3 (2021) : au dessus de la moyenne à partir mi-avril (fin du couvre-feu)

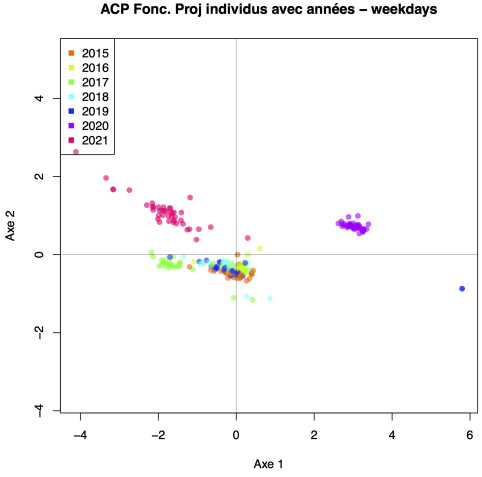


Analyse de transition - Semaines complètes



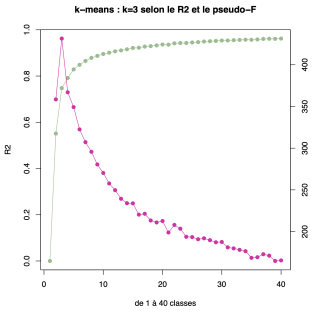
Projections des individus - Semaines hors week-ends

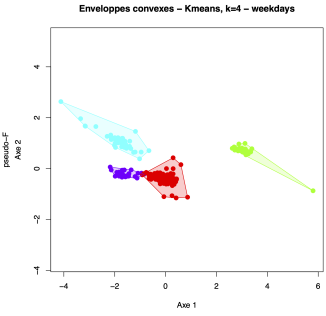
1e plan factoriel (+84% iner.) : à présent, 2017 se démarque également !

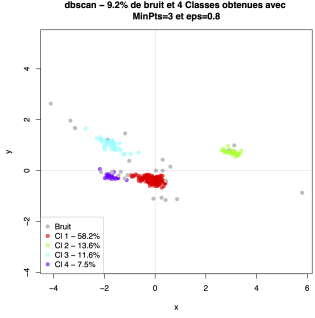


Etudes des clusters - Semaines hors week-ends

- la recherche du nombre de clusters par les critères classiques conduit à 3
- des données de 2017 se concentrent dans la partie centrale gauche
- nous paramétrons les algorithmes pour la construction de 4 classes

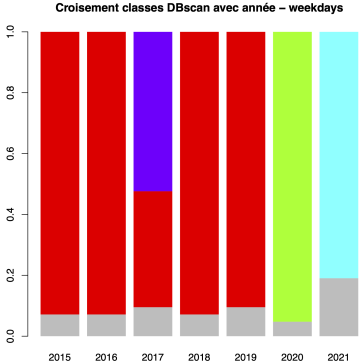






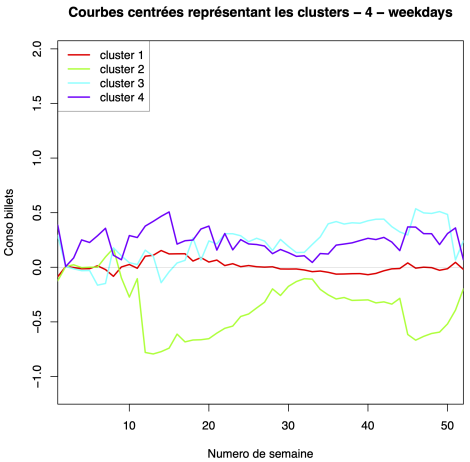
Répartition des clusters - Semaines hors week-ends

Quelle que soit la méthode retenue, les profils des années 2017, 2020 et 2021 se concentrent dans des classes distinctes.

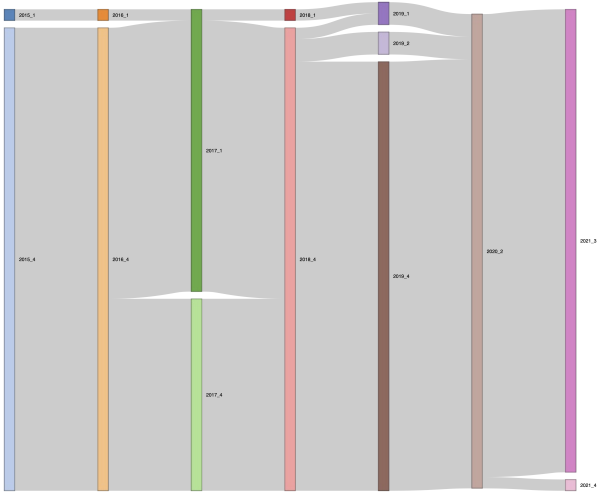


Variations des données d'intensité - Semaines hors week-ends

- cl1 : peu d'écarts à la moyenne
- cl2 (2020) : forte baisse au confinement 1 puis progression jusqu'au confinement 2 et reprise lente en fin d'année
- cl3 (2021) : au dessus de la moyenne à partir mi-avril (fin couvre-feu)
- cl4 (2017) : au dessus de la moyenne régulièrement

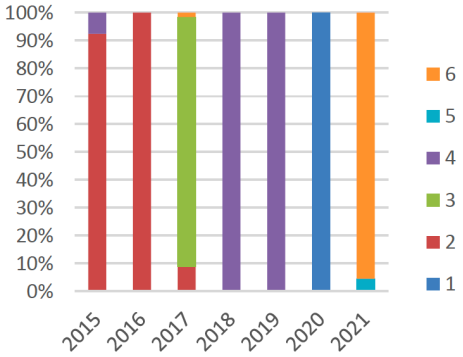


Analyse de transition - Semaines hors weekends



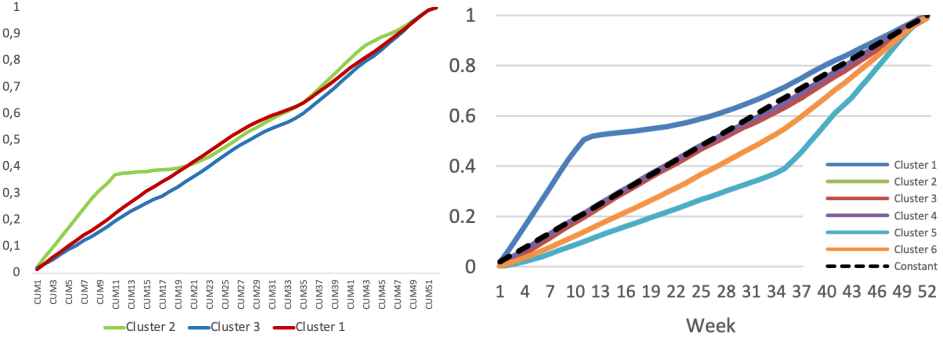
Comparaison avec Montréal - Semaines complètes

Avec la méthode des K-means, nous obtenons des résultats similaires dans les deux métropoles - partition en 3 classes à Lyon (gauche) et 4 classes à Montréal (droite) (Bourdeau & Morency, 2022).



Comparaison avec Montréal - Semaines complètes

L'analyse de la moyenne des validations cumulées montre une plus forte validation en début d'année 2020 avant le confinement - à Lyon (gauche) et à Montréal (droite) (Bourdeau & Morency, 2022).



Conclusion

Principaux enseignements

- usage du réseau de métro, à un niveau agrégé, qui reste stable en structure, malgré une demande croissante
- diminution marquée de l'intensité des validations lors de la crise sanitaire et des mesures restrictives du gouvernement
- grande variabilité de l'intensité des validations entre les stations, suite à la crise, qu'il faudra expliquer
- conclusions à modérer selon la période (jours de semaine ou semaine complète)

Perspectives

- intégrer le cumul hebdomadaire des validations dans l'analyse fonctionnelle et la classification
- tester la pertinence d'algorithmes de partitionnement "flous" (Fuzzy) ou permettant des clusters "imbriqués" (Optics)
- analyser les données relatives aux autres modes de transport en commun (tram, bus à haut niveau de service et bus de ville)
- représenter la distribution spatiale des clusters et mettre en évidence les interactions entre les modes
- poursuivre les comparaisons avec Montréal

Bibliographie I

- Borkowski, P., Jażdżewska-Gutta, M., Szmelter-Jarosz, A. (2021) *Lockdowned : Everyday mobility changes in response to COVID-19*, Journal of Transport Geography, 90, issue C.
- Bourdeau, J.S., Morency, C. (2022) *Yearly patterns of transit usage : a cumulative clustering approach using 7 years of smart card data*, 12th ISCTSC, Lisbon, Portugal.
- Deschaintres, E., Morency, C., Trépanier, M. (2019) *Analyzing Transit User Behavior with 51 Weeks of Smart Card Data*, Transportation Research Record, 2673(6), 33-45.
- Egu, O., Bonnel, P. (2020) *Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. A case study in Lyon*, Travel Behaviour and Society, 19, 112-123.

Bibliographie I

- Guelton S., Poinsot P. (2020) *Mobilités urbaines : quels modèles de financement ?*, L'économie politique, 1(85), 36-46.
- Pelletier, M.-P., Trépanier, M., Morency, C. (2011) *Smart card data use in public transit : A literature review*, Transportation Research Part C : Emerging Technologies, 19(4), 557-568.
- Ramsay, J. O., Silverman, B. W. (2005) *Functional Data Analysis*, Springer.