Peut-on attraper les utilisateurs de Vélo'v au *Lasso* ?

Gabriel MICHAU[1], Céline ROBARDET[2], Luc MERCHEZ[3], Pablo JENSEN[1], Patrice ABRY[1], Patrick FLANDRIN[1], Pierre BORGNAT[1]

[1]Laboratoire de Physique (UMR CNRS 5672), École normale supérieure de Lyon, Université de Lyon, CNRS
46, allée d'Italie, 69364 Lyon cedex 07, France

[2]Université de Lyon, INSA-Lyon, CNRS, LIRIS (UMR CNRS 5205);
Batiment Blaise Pascal 69621 Villeurbanne cedex, France

[3]Équipe BIOGÉOPHILE, UMR CNRS 5600 « Environnement Ville Société », École normale supérieure de Lyon,
Université de Lyon, 15 parvis René Descartes, BP 7000 69342 Lyon cedex 07, France

gabriel.michau@ens-lyon.fr; celine.robardet@insa-lyon.fr; luc.merchez@ens-lyon.fr;
pablo.jensen@ens-lyon.fr ; patrice.abry@ens-lyon.fr; patrick.flandrin@ens-lyon.fr;
pierre.borgnat@ens-lyon.fr

Résumé – Nous étudions les statistiques des déplacements en Vélo'v, un système public et automatisé de location de vélos à Lyon, en essayant de trouver un modèle statistique de régression reliant les données socio-économiques décrivant les populations des quartiers aux flux de vélos entrants et sortants. Pour cela, on s'intéresse à la régression linéaire parcimonieuse avec contraintes des positivités : la solution du *lasso* n'étant pas satisfaisante quand on tente de l'interpréter, on étudie le problème de l'estimation en présence d'observations tronquées et on propose un algorithme pour adapter le résultat du LARS dans ce cas. Nous étudions la méthode proposée pour montrer qu'elle fonctionne sur des cas simulés et retrouve le modèle linéaire des demandes tronquées dans les observations.

Abstract – A statistical analysis of the trips made using Velo'v, the community shared bicycle system in Lyon city, is conducted in attempt to relate, through a statistical regression model, social, demographic and economical data of the various neighborhoods of the city with the actual trips made from and to the different parts of the city. For that, parcimonious linear regression methods with positivity constraints are studied. As the *lasso* solution is found to be not fully adequate for the present problem when interpretation is attempted, a further specificy of he data is introduced: real data on demand are truncated because of a constraint of capacity in bikes at the stands. The LARS algorithm is adapted to this situation of truncated observations. A study of the proposed new algorithm is conducted and it is found to perform well in controlled simulations, estimating well the linear model used to generate truncated observations.

1 Système Vélo'v : comment analyser l'usage qu'en fait par la population ?

Nous étudions le service Vélo'v de mise à disposition de vélos en tant que transport public. Ce système automatisé de location de vélos est en service à Lyon depuis 2005. Sa dynamique de fonctionnement s'apparente par de multiples aspects à celle d'un système complexe, c'est-à-dire qu'au-dessus de comportements individuels simples, on obtient un comportement collectif complexe.

Dans des premiers travaux [1, 2], nous avons formulé des analyses globales de la dynamique temporelle du service et de la répartition spatiale des trajets. Le présent travail est d'étudier des méthodes statistiques pour relier les données socio-économiques de la ville (venant de l'INSEE) avec l'usage qu'il est fait des Vélo'V, quartier par quartier et dépendant du moment de la journée.

Pré-traitement et formalisation. Le problème est restreint dans ce travail à la recherche d'un modèle supposé linéaire (à défaut d'autres a priori) entre les flux de Vélo'v aux stations et les caractéristiques démographiques et socio-économiques du quartier entourant chaque station.

Le premier point est de réconcilier les niveaux de description différents entre les données : les flux de Vélo'v sont connus à chaque station, tandis que les variables démographiques et socio-économique sont connues à l'échelle des ilots (morceau de quartier) ou des IRIS (Ilots Regroupés pour l'Inférence Statistique, découpés par l'INSEE) à l'échelle au-dessus. Les stations sont souvent sur ou près des axes majeurs de circulations qui découpent aussi la ville en IRIS : associer les stations à l'IRIS le plus proche n'est clairement pas une solution acceptable. Inspirés alors par l'utilisation des modes *raster* en géographie, il est logique de lisser les flux de Vélo'v et de les réaffecter aux IRIS par un noyau de lissage, tout en conservant le nombre total de trajets effectués. Pour faire cela, notons $\#v(k,t)$ un flux de Vélo'v à la station $k \in \mathcal{S}$ entre le temps t et $t+\Delta$ (avec typiquement $\Delta = 1$h en accord avec [1]). On utilise un lissage spatial local de longueur typique R_0 qui correspond à ce mode *raster* ou à proposer une hypothèse sur un noyau de

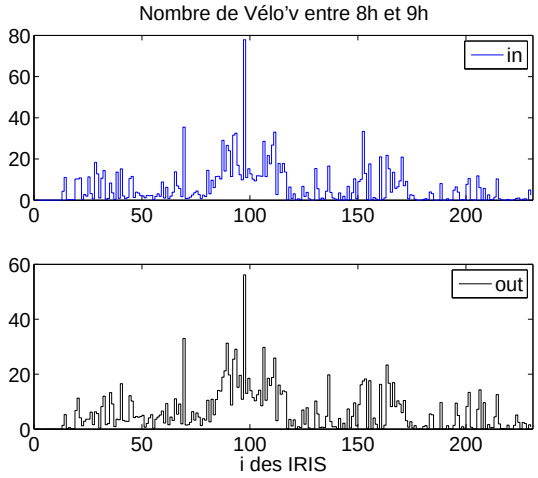

FIG. 1 – Nombres de Vélo'v sortant et arrivant aux IRIS le matin les jours de semaine entre 8h et 9h, en moyenne sur la période du 11/2005 au 05/2006.

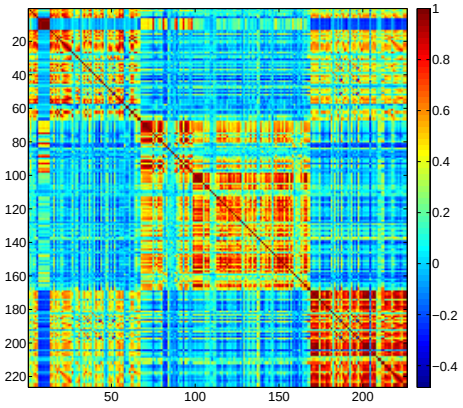

FIG. 2 – Matrice de corrélation des 227 données INSEE disponibles. Les fortes corrélations sont en rouge, anti-corrélations en bleu foncé et les corrélations faibles sont en bleu ciel ou vert.

la probabilité qu'un utilisateur (ou toutes autre donnée INSEE) de l'IRIS i influe sur les trajets à la station k ; proposer une densité exponentielle donne un terme en e^{-r_{ik}/R_0} où r_{ik} est la distance entre le centre de l'IRIS i et la station k. En normant la probabilité, on estime le flux par IRIS comme il suit :

$$F_i(t) = \sum_{k \in S} \#v(k,t) \cdot \left(e^{-r_{ik}/R_0} / \sum_{i \in \{\text{IRIS}\}} e^{-r_{ik}/R_0} \right). \quad (1)$$

Cette réaffectation des déplacements en Vélo'v aux IRIS conserve les flux de Vélo'v :

$$\sum_{i \in \{\text{IRIS}\}} F_i(t) = \sum_{k \in \mathcal{S}} \#v(k,t). \quad (2)$$

L'hypothèse implicite est que les utilisateurs font aisément une distance de R_0 entre leur départ (ou destination) et la station de Vélo'v (typiquement, $R_0 = 100$ m). Les flux de Vélo'v entrants F^{in} et sortants F^{in} pour chaque IRIS entre 8h et 9h du matin sont représentés sur la Fig. 1. On posera que n est le nombre d'IRIS étudiés sur Lyon, ainsi $F^{in/out} \in {\mathbb{R}_+}^n$. Sur l'ensemble de la ville, $n = 230$. Les flux entrants et sortants le matin ne sont pas les mêmes partout : certaines zones se vident tandis que d'autres se remplissent (ceci fut discuté dans [2]). Un objectif principal de l'étude est donc de savoir modéliser ce qui explique cette différence de comportements en le reliant aux caractéristiques démographiques et socio-économiques des quartiers de la ville.

Les variables démographiques et socio-économiques, positives et en nombre p, sont bien souvent corrélées les unes avec les autres, ainsi qu'illustré en Fig. 2. La matrice de ces variables par IRIS, normées mais non centrées, notée $X \in \mathbb{R}^{n \times p}$, est ainsi mal conditionnée. Un deuxième pré-traitement sera de limiter ici l'étude à une sélection d'un petit nombre de variables $p = 25$, sélectionnés avec des économistes du transport[1]. Les données de régression et les données de flux de Vélo'v à modéliser sont toutes positives. Il n'est pas évident que l'on doive chercher des coefficients de régression positifs. Cependant, si l'on veut pouvoir comparer cette étude à des analyse par des modèles-agents qui est souvent la norme dans ces études de systèmes complexes, on est conduit à restreindre les modèles utiles à ceux dont les coefficients sont positifs : en effet, des coefficients positifs privilégient les explications par motifs de déplacements (la présence de tels types de population ou de tels types d'activités explique des trajets) tandis que des coefficients négatifs correspondraient à des effets inhibiteurs que l'on a plus de mal à maîtriser et inclure dans les modèles d'agents.

Le problème que l'on propose revient donc à chercher les coefficients de régression $\beta^{in/out}(t)$ tels que

$$F^{in/out}(t) \simeq X\beta^{in/out}(t) \quad \text{et} \quad \beta^{in/out} \geq 0 \quad (3)$$

dans la situation où $p < n$.

Solution du *lasso*. Ne sachant pas a priori quelles données sont importantes, on voudrait trouver un modèle parcimonieux, plus aisé à interpréter. Des approches de régression simples : moindres carrés ordinaires ou partiels, régression *ridge*, trouvent toujours dans la masse de données des corrélations apparentes entre variables et flux, peu interprétables. Nous privilégions donc une approche par régression parcimonieuse avec un problème de type *lasso* positif [3, 4] :

$$\hat{\beta}^{in/out}[s] = \arg\min_{b \geq 0} \| F^{in/out} - Xb \|_2 \quad \text{t.q.} \quad \| b \|_1 \leq s. \quad (4)$$

La solution obtenue (à un temps t) par l'algorithme du LARS [4] est montré en Fig. 3 pour l'évolution des solutions (en fonction de la norme de la contrainte s, normalisée entre 0 et 1) et en Fig. 4 pour la comparaison de l'ajustement obtenu avec des données mesurées.

Critique du modèle supposé et de la solution. Le s optimal pourrait être décidé par validation croisée. Ici, supposer une contrainte de positivité et ne pas avoir centré les variables suffit à sélectionner un sous-ensemble de variables, 4 dans $\hat{\beta}^{in}$ et 6 dans $\hat{\beta}^{out}$, même pour s maximal (normalisé à 1). Le pro-

[1] Nous remercions Alain Bonnafous du LET (UMR CNRS 55593 et Université Lyon 2) pour les discussions fructueuses à ce sujet.

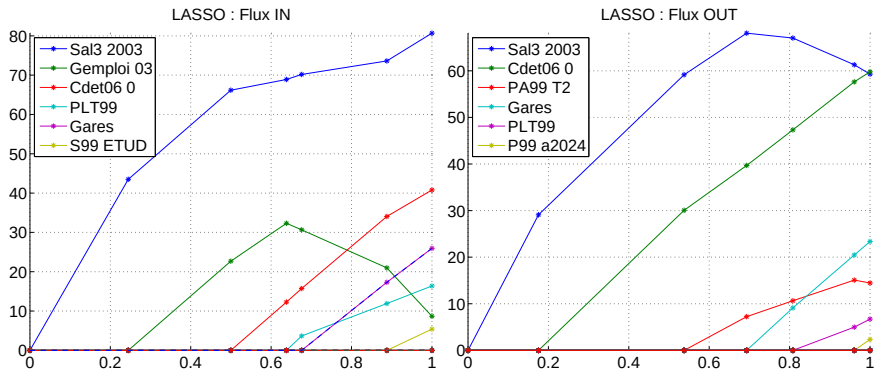

FIG. 3 – Estimation des coefficients β du *lasso* positif, fonction de s, pour $t = 8$h. Les noms des variables retenues sont en légende dans l'ordre d'apparition.

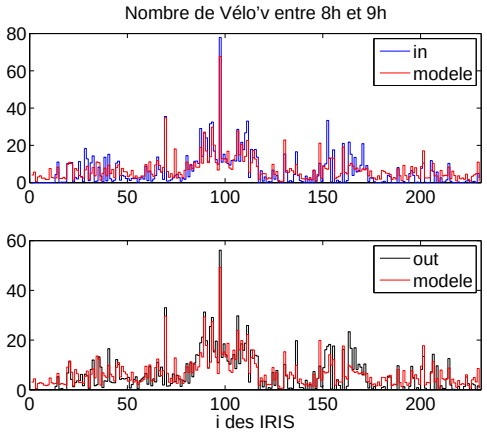

FIG. 4 – Estimations par *lasso* positif (en rouge) et données (en bleu ou noir), pour $t = 8$h

blème est que les variables obtenues sont essentiellement les mêmes, qu'on regarde les flux entrants ou sortants (et ce serait de même matin ou soir). Les variables sont principalement les suivantes : `Sal3 2003` (nombre salariés du secteur tertiaire), `Gemploi03` (emplois dans l'IRIS), `Cdet06 0` (nombre de commerces), `PLT99` (population au lieu de travail), `PA99T2` (population déclarer se déplacer à pieds), `S99ETUD` ou `P99 a2024` (population étudiante ou d'âge 20-24 ans), `Gares` (présence d'une gare dans l'IRIS). Bien que ceci renseigne sur qui utilise les Vélo'v (population se déplaçant à pieds, étudiants,...) et pour quoi (arriver sur des lieux de travail, de commerce, les gares ou leurs habitations), on n'arrive pas par cette méthode à trouver de différences entre les usages du matin et du soir qu'on sait (par les enquêtes terrain auprès de quelques usagers) être différents (plutôt pour aller au travail le matin, en revenir le soir). Il manque un élément au problème et nous proposons une méthode pour corriger le *lasso* pour le cas étudié.

2 Revisiter la régression parcimonieuse avec observation tronquée

Formulation du problème. En réalité, le système fonctionne avec une contrainte supplémentaire qui change le programme de régression à résoudre : la contrainte de capacité des stations. La demande réelle qu'on voudrait estimer par le modèle linéaire, est parfois tronquée du fait que les stations sont facilement pleines ou vides à certaines heures, ainsi qu'étudié dans [5]. Cela réduit les trajets respectivement entrants ou sortants, par rapport à la demande. Notons $C \in \mathbb{R}^n$ la capacité en bornes de Vélo'v d'un IRIS, et $N^0(t)$ le nombre de Vélo'v présents initialement. On a toujours $N_i^0 \in [0,\ C]^n$ et pour les flux, la contrainte est :

$$F^{in}(t) - F^{out}(t) + N^0(t) \in [0,\ C]^n. \tag{5}$$

En réalité donc, un meilleur modèle génératif des flux passe par donner un modèle de demande $D^{in/out}$, avec erreurs éventuelles $\epsilon^{in/out}$, qui est ensuite tronquée éventuellement par la capacité des stations en vélos ou en places libres :

$$\begin{aligned} &D^{in} = X\beta^{in} + \epsilon^{in} \\ &D^{out} = X\beta^{out} + \epsilon^{out} \\ &\text{avec } \begin{cases} F^{in} = \min(D^{in}, F^{out} + C - N^0) \\ F^{out} = \min(D^{out}, F^{in} + N^0). \end{cases} \end{aligned} \tag{6}$$

Régression parcimonieuse avec observation tronquée. La contribution méthodologique de ce travail est de s'intéresser à l'estimation des coefficients de régression dans ce problème non standard, sans perdre la parcimonie des solutions. Pour cela, nous proposons de s'intéresser à l'algorithme qui suit et de le tester dans un premier temps sur des données simulées. On suppose qu'un estimation moyenne $\hat{N}^0$ de N^0 est connue (ce qu'on a par des relevés pour les données Vélo'v, ainsi que discuté dans [5]).

- On procède itérativement en augmentant l'ensemble $\mathcal{I}^{\mathcal{T}}$ où les demandes sont peut-être tronquées. A l'étape initiale, $k = 0$ et $\mathcal{I}^{\mathcal{T}} = \varnothing$ et on retient les solutions $\hat{\beta}^{in/out}(k = 0)$ du *lasso* positif, eq.(4). Certains IRIS auront un mauvais ajustement car la demande est plus grande que le flux réel à cause de la troncature par capacité.
- A l'étape k on ajoute l'IRIS i_k dans $\mathcal{I}^{\mathcal{T}}$ comme candidat où la demande est tronquée, obtenu ainsi :

$$\begin{aligned} i_k = \quad \arg\min_i \min(&|(F^{in} - F^{out} + \hat{N}^0 - C)_i|, \\ &|(F^{in} - F^{out} + \hat{N}^0)_i|) \end{aligned} \tag{7}$$

- On trouve un ajustement parcimonieux positif par eq.(4) pour les IRIS de l'ensemble complémentaire $\bar{\mathcal{I}}^{\mathcal{T}}$ de $\mathcal{I}^{\mathcal{T}}$:

$$\begin{cases} \hat{\beta}^{in}[s](k) = \arg\min_{b\geq 0} \parallel (F^{in} - Xb)_{\bar{\mathcal{I}}^{\mathcal{T}}} \parallel_2 \\ \qquad\qquad \text{t.q.} \quad \parallel b \parallel_1 \leq s \\ \hat{\beta}^{out}[s](k) = \arg\min_{b\geq 0} \parallel (F^{out} - Xb)_{\bar{\mathcal{I}}^{\mathcal{T}}} \parallel_2 \\ \qquad\qquad \text{t.q.} \quad \parallel b \parallel_1 \leq s \end{cases} \tag{8}$$

  La parcimonie de la solution est contrôlée par le choix de s.
- On calcule les erreurs obtenues sur les complémentaire de

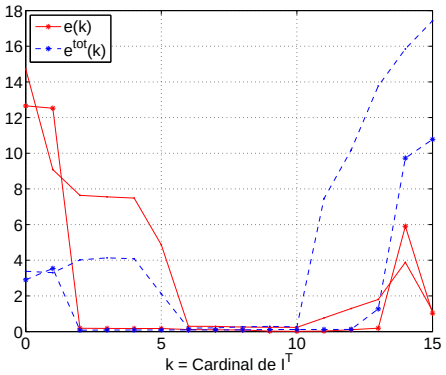

FIG. 5 – Erreur $e(k)$ et erreurs d'oracle ($D - X\hat{\beta}$) (courbes avec * pour in, · pour out)). $k = 6$ est l'optimum.

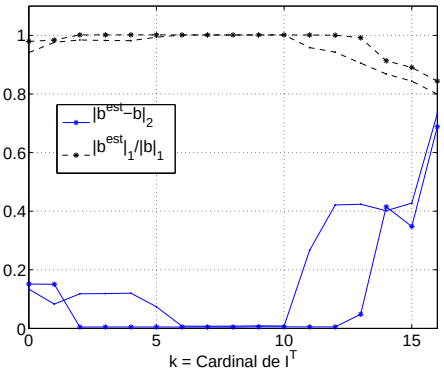

FIG. 6 – Erreur en ℓ_2 sur β et comparaison des norme ℓ_1 de $\hat{\beta}$ et β vrais.

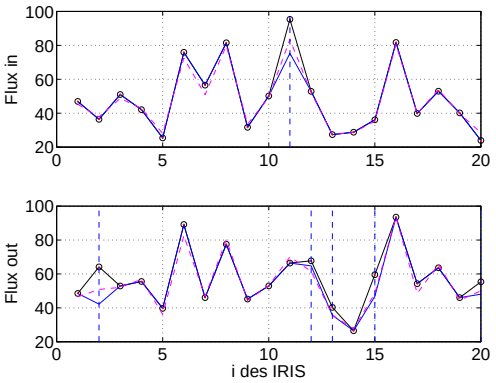

FIG. 7 – Estimations par *lasso* sans (pointillés) ou avec correction pour toncature (points rouges) des demandes (noir) ou flux (bleu).

$\mathcal{I}^{\mathcal{T}}$, à la fois pour les flux entrants et sortants :

$$\begin{aligned} e^{in}(k) &= \| (F^{in} - X\hat{\beta}^{in}[s](k))_{\bar{\mathcal{I}^{\mathcal{T}}}} \|_2 \\ e^{out}(k) &= \| (F^{out} - X\hat{\beta}^{in}[s](k))_{\bar{\mathcal{I}^{\mathcal{T}}}} \|_2 \end{aligned} \quad (9)$$

Génériquement, ces deux erreurs vont d'abord diminuer tant qu'on ajoute dans $\mathcal{I}^{\mathcal{T}}$ des stations dont la demande est tronquée (car enlever ces stations permet un meilleur ajustement), puis elles vont réaugmenter lentement si on enlève d'autres (car enlever les autres réduit le nombre d'observations disponibles). Après avoir ajouté toutes les observations (IRIS ici), le choix de l'optimum de k à retenir est donné par le premier k où $e^{in}(k)$ et $e^{out}(k)$ sont passés par un premier minimum local.

3 Test de la méthode

Test sur des données simulées. L'application de cette méthode est évaluée par simulations numériques. On utilise l'éq. (6) comme modèle génératif, avec des paramètres dimensionnés pour le cas Vélo'v. On prend $n = 20, p = 10$. Les capacités C sont i.i.d uniformes dans $[5,\ 35]$, les $(N^0)_i$ uniformes dans C_i. On utilise des fausses variables X de lois i.i.d. uniforme dans $[0,\ 40]$, des coefficients $\beta^{in/out}$ dans $[0,\ 1]$, parcimonieux (de $\| \beta \|_0 \leq 6$). On ajoute une erreur aléatoire i.i.d. $\epsilon \sim \mathcal{N}(0, 1)$ aux demandes obtenues, pour s'approcher de conditions réalistes d'utilisation.

Les figures 5 à 7 montrent les résultats de la méthode décrite : décroissance de $e^{in}(k)$ et $e^{out}(k)$ à gauche, comparée aux erreurs d'un oracle ($D - X\hat{\beta}$) ; erreur obtenus sur les coefficients estimés $\hat{\beta}$ et norme ℓ_1 de l'estimation (normalisé par $\| \beta \|_1$). On voit qu'ici, $k = 6$ est l'optimum à retenir. La dernière figure 7 compare les donnés simulées à ces estimations : les demandes sont en noir, les flux tronqués observés en rouge (avec 5 endroits où la limite est active dans cet exemple). La méthode proposée retrouve bien ces endroits et des estimations convenables là où le *lasso* non corrigé ne le pouvait pas.

4 Conclusion

Le résultat de ce travail est, en comparaison de ce que donnent les algorithmes usuels de régression parcimonieuse, on est capable de donner une meilleure estimation parcimonieuse des coefficients d'un modèle avec une observation tronquée, ici du fait d'une contrainte de capacité,

Les résultats sur des données réelles, à savoir le problème initial de modélisation des demandes en trajets en Vélo'v et leurs relations avec les populations, emplois, commerces, etc. présents sur les lieux, feront l'objet de travaux futurs. Une autre perspective méthodlogique du travail est, dans le futur, de trouver à inclure l'étape de sélection entre des variables très corrélées dans l'estimation du modèle, puisque l'on sait que le *lasso* sur lequel se base la méthode, se comporte mal pour des variables mal conditionnées.

Références

[1] P. Borgnat, P. Abry, P. Flandrin. Symposium GRETSI-09, Dijon, FR (Sept., 2009).

[2] P. Borgnat, C. Robardet, J.-B. Rouquier, P. Abry, E. Fleury, et P. Flandrin. *Advances in Complex Systemes*, à paraître (2011).

[3] R. Tibshirani. *J. Roy. Statist. Soc. Ser. B*, **58**, p. 267–288 (1996).

[4] B. Efron, T. Hastie, I. Johnstone et R. Tibshirani. *The Annals of Statistics*, **32** :2, p. 407–499 (2004).

[5] L. Merchez et J.-B. Rouquier. "L'usage des vélos en libre service (VLS) comme révélateur des rythmes urbains : le cas des stations de Vélo'v à Lyon", soumis à *Données Urbaines* (2010).