Prévision de la fréquentation d'un réseau de transport à l'aide de modèles additifs généralisés

Léna Carel [1,2] & Pierre Alquier [2]

[1] *TRANSDEV GROUP*
3 allée de Grenelle - CS 20098, 92442 Issy-les-Moulineaux Cedex - France
[2] *CREST, ENSAE, Université Paris-Saclay*
5 avenue Henry Le Chatelier, 91120 Palaiseau - France
lena.carel@ensae.fr, pierre.alquier@ensae.fr

Résumé. Afin de planifier au mieux les services des différentes lignes d'un réseau de transport, il est important de bien appréhender la demande des usagers de ce réseau, pour leur offrir un service optimal. Dans cet article, nous proposons donc d'utiliser les modèles additifs généralisés (GAM) pour prévoir par heure et par arrêt la fréquentation du réseau.

Mots-clés. Analyse de données, fouille de données, Apprentissage et classification, Environnement, climat, Villes intelligentes.

Abstract. In order to improve the quality of an urban transportation network, it is very important to be able to anticipate the users' demand. In this paper, we propose to use a General Additive Model (GAM) to predict the number of passengers at a tramway stop hour by hour.

Keywords. Data analysis, data mining, Machine learning and clustering, Environment, climate, Smart cities.

1 Introduction

Avec l'arrivée grandissante de capteurs et autres systèmes de collecte de données dans nos environnements proches, l'informatique et les systèmes numériques sont des outils de plus en plus présents dans l'aide à la prise de décision et la planification urbaine [9]. Par exemple, dans le domaine des transports, on trouve une littérature florissante sur l'utilisation et l'analyse des données de validation billetique dans les transports en commun [5], les systèmes de partage de vélos [8] ou bien de taxi [6].

Les modèles GAM [4] sont très populaires en machine learning: en effet, d'une part, ils sont assez flexibles pour estimer des effets non linéaires, et d'autre part, leur complexité n'explose pas de façon exponentielle avec le nombre de variables, permettant d'éviter le fléau de la dimension. Ces modèles ont été utilisés avec succès dans de nombreux domaines, cf. par exemple la prévision à court terme de la charge du réseau électrique [7]. Nous allons nous en servir ici pour prévoir la fréquentation du réseau Transdev de Rouen.

2 Données

Nous avons à notre disposition les données de validation billetique du réseau Astuce de la métropole Rouen-Normandie sur un historique de deux ans (1^{er} janvier 2014 au 31 décembre 2015). Pour pouvoir prédire la fréquentation par arrêt, nous avons agrégé les données par jour, heure, arrêt et ligne. Un arrêt commercial est divisé en plusieurs arrêts physiques en fonction notamment du sens de la/des ligne(s) qui y passent, nous avons un total de 2308 arrêts physiques pour 615 arrêts commerciaux. Dans un soucis de clarté, nous indiquons ici au lecteur que lorsque nous parlons d'arrêt, nous faisons allusion à un arrêt physique. Ainsi, nous obtenons une base de données contenant 8 198 346 observations.

De part la taille des données, nous ne modéliserons ici que la fréquentation d'un seul arrêt. Nous avons choisi l'arrêt générant le plus de trafic, qui est l'arrêt de Métro « Théâtre des Arts » dans la direction du nord vers le sud.

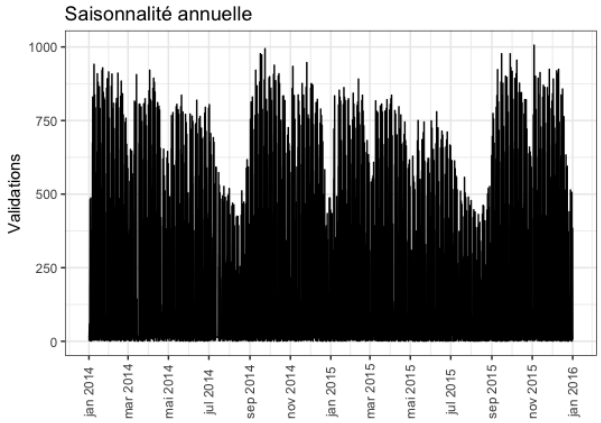
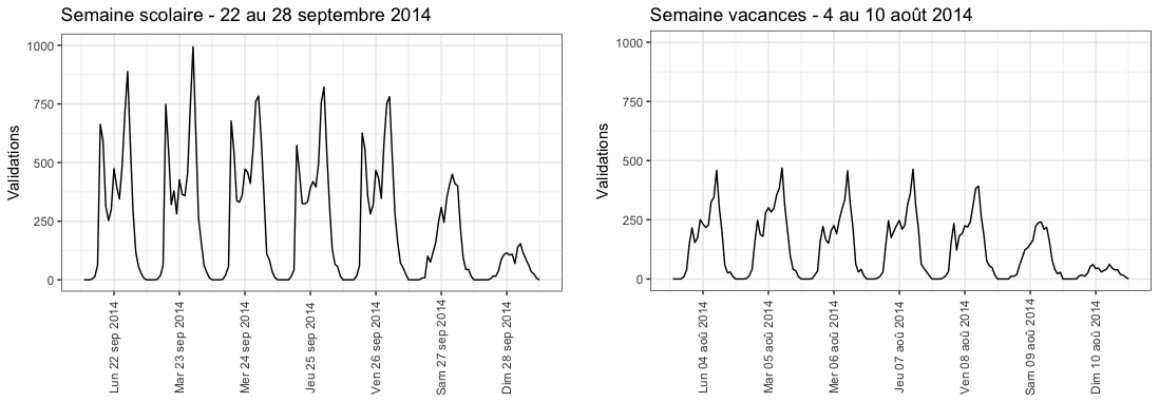


FIG. 1 – *Fréquentation de l'arrêt «Théâtre des Arts» entre le 1er janvier 2014 et le 31 décembre 2015*

En regardant l'évolution annuelle de la fréquentation à l'arrêt (Figure 1), on note que les fréquentations semblent être plus nombreuses en septembre puis diminuent tout au long de l'année scolaire. De plus, on remarque que pendant les périodes de vacances scolaires, les validations sont beaucoup moins nombreuses.

En effet, en observant la Figure 2 on note clairement la différence de fréquentation entre la semaine scolaire et la semaine de vacances. On remarque aussi que les profils journaliers diffèrent entre ces deux types de périodes. En période scolaire, on observe des pics de fréquentation le matin et le soir en semaine, ainsi qu'un léger pic à midi. En période de vacances, on remarque un pic en début de soirée alors que le matin et le midi ces pics sont moins prononcés qu'en période scolaire. Cependant, pour les deux périodes on note qu'il y a des fréquentations à peu près aussi élevées qu'en heure creuses le samedi, et très peu de validations le dimanche. Ainsi, on ajoute aux données plusieurs variables



(a) *Semaine scolaire* (b) *Semaine de vacances*

FIG. 2 – *Comparaison de la fréquentation de l'arrêt « Théâtre des Arts» entre une semaine de période scolaire et une semaine de vacances*

TAB. 1 – *Description des variables*

Données calendaires		Données météorologiques	
Variable	Description	Variable	Description
`Heure`	Heure de la journée	`t_moy`	Température moyenne de la journée
`nMois`	Numéro du mois	`ff_moy`	Vitesse moyenne du vent de la journée
`nJourSem`	Numéro du jour dans la semaine	`u_moy`	Humidité moyenne de la journée
`nJour`	Numéro du jour dans l'année	`rr`	Nombre de millimètres de précipitation de la journée

calendaires permettant de prendre en compte toutes ces saisonnalités. De plus, il nous semble important d'enrichir les données à l'aide de données météorologiques [1]. L'ensemble de ces variables sont décrites dans la Table 1.

Afin de vérifier la performance de nos modèles, et pour éviter le sur-apprentissage, nous allons séparer nos données en deux bases distinctes. Nous garderons d'un côté les données du 1^{er} janvier 2014 au 30 novembre 2015 pour former notre base d'apprentissage. Les données du mois de décembre 2015 serviront quant à elles de base de test pour les modèles.

1. base SYNOP de Météo-France

3 Méthodologie

3.1 Modèle

Le modèle linéaire, populaire en régression, repose sur l'hypothèse d'un lien linéaire entre les prédicteurs et la variable à prédire. Le lien entre certaines variables, comme la température et le nombre de passagers à un arrêt de bus, a ceci dit peu de chance d'être linéaire. On propose donc le modèle GAM:

$$Y_i = \beta_0 + \sum_{j=1}^{p} s_j(X_{i,j}) + \epsilon_i. \tag{1}$$

De cette manière, on remplace les coefficients de régression β_j par des fonctions $s_j(\cdot)$. On peut choisir différentes méthodes d'estimation des fonctions s_j inconnues, nous avons choisi d'utiliser les splines cubiques implémentées par défaut dans la fonction `gam` du package `mgcv` en `R`.

3.2 Intervalle de confiance

Dans le but d'adapter le nombre de bus à la demande, une société de transport ne sera pas intéressé par une prédiction ponctuelle. Il faut plutôt savoir combien de passagers au minimum et au maximum pourront être présents à un arrêt à un instant donné. Autrement dit, il faut absolment accompagner la prévision d'un intervalle de confiance. Or dans le modèle GAM (1), il semble naturel que l'incertitude de la prédiction dépende des différentes variables. En effet, l'incertitude aux heures de pointes dépend des conditions de trafic par exemple et sera bien plus grande que l'incertitude aux heures creuses où on sait que le nombre de passagers est faible quoi qu'il arrive. On suppose donc que $\epsilon_i \sim N(0,\sigma_i^2)$ et on propose

$$\mathbb{E}[\epsilon_i^2] = \gamma_0 + \sum_{j=1}^{p} t_j(X_{i,j}).$$

On estime alors l'intervalle de confiance à 95% comme $Y_i \in [\hat{Y}_i - 1.96\ \hat{\sigma}_i^2\ ;\ \hat{Y}_i + 1.96\ \hat{\sigma}_i^2]$. Si cette modélisation est possible avec le package `mgcv`, nous avons préféré effectuer le travail d'analyse sur les ϵ_i à la main. En effet, ce travail de modélisation nous semble essentiel: nous souhaitons dans une version approfondie de cet article étudier la persistance de la variance de l'erreur, via un modèle inspiré des ARCH ou GARCH en finance. Par exemple, on peut proposer:

$$\mathbb{E}[\epsilon_i^2] = \gamma_0 + \gamma_1 \epsilon_{i-1}^2 + \sum_{j=1}^{p} t_j(X_{i,j}).$$

4 Choix du modèle

Nous avons entraîné différents modèles sur l'échantillon d'apprentissage et les avons testés sur l'échantillon de test. Le modèle retenu est composé de six sous-modèles: un pour les jours fériés, un pour les jours de pont, un pour les samedis, un pour les dimanches, un pour les lundis à vendredis en période de vacances et un pour les lundis à vendredis en période scolaire. La Figure 3 représente la prévision de la fréquentation ainsi que de l'intervalle de confiance sur une partie de l'échantillon de test. Nous avons selectionné la période du 14 au 27 décembre 2015, car elle contient une semaine de période scolaire, une semaine de vacances et un jour férié (25 décembre 2015).

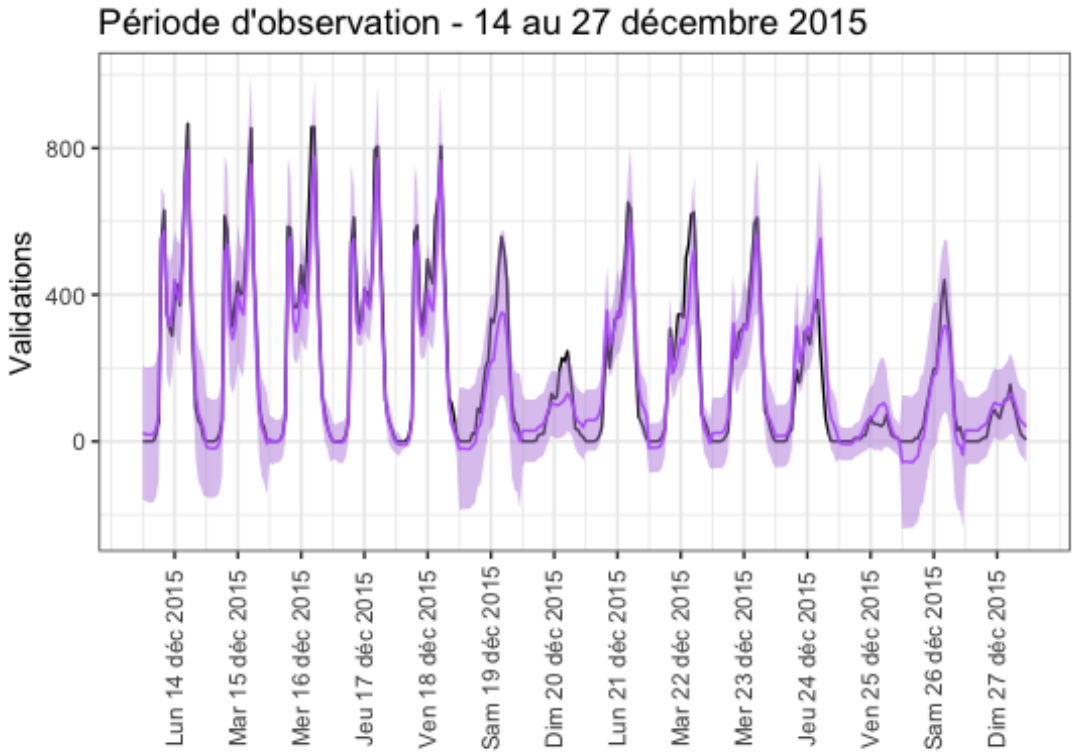


FIG. 3 – *Prévision de la fréquentation à l'arrêt «Théâtre des Arts»*

En observant la prévision obtenue sur notre période, on remarque que la fréquentation observée est globalement contenue dans l'intervalle de confiance estimé. On note cependant que quelques points se situent au-dessus de l'intervalle de confiance, notamment le dimanche 20 décembre. Il faut noter ceci dit que la période des fêtes de Noël est vraisemblablement atypique et n'est représentée qu'une fois dans la base d'apprentissage. Ces premiers résultats sont bien entendu illustratifs, une étude plus systématique sera menée dans un article à venir.

5 Conclusion

Dans ce travail nous avons utilisé les modèles additifs généralisés pour prévoir la fréquentation d'une station sur le réseau de transport de Rouen. Nous avons mis en lumière le fait que pour prévoir au mieux cette fréquentation, il fallait traiter différemment les jours si ils sont fériés, samedi ou dimanche. Pour étendre cette étude à l'intégralité du réseau, il faudrait réaliser un couple de modèles pour chaque arrêt physique.

Si les premiers résultats sont encourageants, il convient bien entendu de mener une étude systématique des modèles possible, pour Y_i et pour ϵ_i afin de déterminer le modèle le plus adapté non seulement pour la prédiction mais aussi pour les intervalles de confiance. On pourra alors comparer la cohérence des modèles obtenus arrêt par arrêt avec les clusters d'arrêts obtenus par les méthodes proposées dans [3, 1, 2]. On pourra enfin étudier la pertinence de nos intervalles de confiance pour détecter des incidents du réseau: afflux inhabituel de passager, pics de fraudes...

Références

[1] L. Carel and P. Alquier. Non-negative matrix factorization as a pre-processing tool for travelers temporal profiles clustering. In M. Verleysen, editor, *Proceedings of the 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 417–422. i6doc.com, 2017.

[2] L. Carel and P. Alquier. Simultaneous Dimension Reduction and Clustering via the NMF-EM Algorithm. *ArXiv e-prints*, Sept. 2017.

[3] M. K. El Mahrsi, E. Côme, J. Baro, and L. Oukhellou. Understanding passenger patterns in public transit through smart card and socioeconomic data: A case study in rennes, france. In *ACM SIGKDD Workshop on Urban Computing*, 2014.

[4] T. Hastie and R. Tibshirani. *Generalized additive models*. Wiley Online Library, 1990.

[5] C. Morency, M. Trepanier, and B. Agard. Measuring transit use variability with smart-card data. *Transport Policy*, 14(3):193–203, 2007.

[6] C. Peng, X. Jin, K.-C. Wong, M. Shi, and P. Liò. Collective human mobility pattern from taxi trips in urban area. *PloS one*, 7(4):e34487, 2012.

[7] A. Pierrot and Y. Goude. Short-term electricity load forecasting with generalized additive models. In *Proceedings of ISAP power*, pages 593–600, 2011.

[8] A. N. Randriamanamihaga, E. Côme, L. Oukhellou, and G. Govaert. Clustering the Vélib' origin-destinations flows by means of Poisson mixture models. In *ESANN*, 2013.

[9] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):38, 2014.