


HMMA 307 : Modèle linéaire avancés

Analyse du nombre de passages de cyclistes à Seattle

Emma Santinelli

 <https://github.com/emmas2210/HMMA307>

Université de Montpellier



Table of Contents

- 1 Les données
- 2 Interprétation des données
- 3 Conclusion

Table des matières

- 1 Les données
- 2 Interprétation des données
- 3 Conclusion

Présentation des données

Les données étudiées sont issues d'un compte effectué, chaque heure, sur le Fremont Bridge à Seattle. Elles sont comptabilisées par un compteur automatique, sensible aux déplacements, installé depuis 2012 sur ce pont.

Les données sont réparties selon :

- 136334 heures d'observation ;
- 48 variables explicatives (les heures).

Constitution du tableau de données

	Total	East	West
Date			
2012-10-03 00:00:00	13.0	4.0	9.0
2012-10-03 01:00:00	10.0	4.0	6.0
2012-10-03 02:00:00	2.0	1.0	1.0
2012-10-03 03:00:00	5.0	2.0	3.0
2012-10-03 04:00:00	7.0	6.0	1.0

Description statistique de ces données :

- 113 vélos en moyenne (51 par l'est et 61 par l'ouest)
- 1097 vélos au maximum (11 Juillet 2019 à 17h)
- 0 vélo minimum (1 er Novembre 2012 à 2h)

Figure: Visualisation simple des données

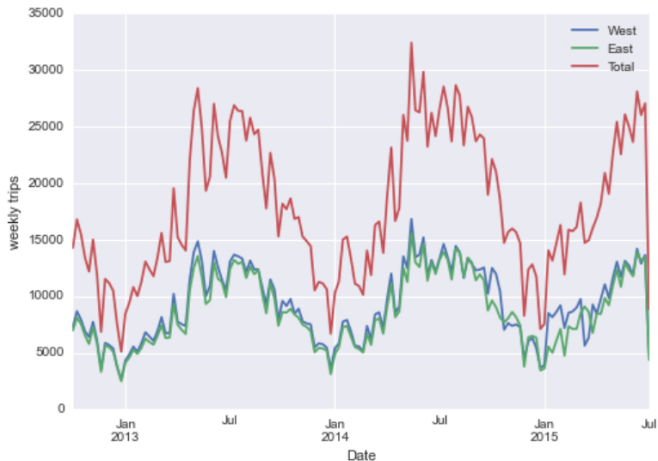


Table des matières

- 1 Les données
- 2 Interprétation des données**
- 3 Conclusion

Transformation des données

Figure: Nouveau tableau de données

	East										...	West											
Date	0	1	2	3	4	5	6	7	8	9	...	14	15	16	17	18	19	20	21	22	23		
2012-10-03	4	4	1	2	6	21	105	257	291	172	...	51	92	182	391	258	69	51	38	25	12		
2012-10-04	7	3	3	0	7	15	91	230	284	147	...	56	74	161	353	241	107	56	39	21	30		
2012-10-05	4	4	4	2	7	18	68	218	251	131	...	62	84	190	290	209	73	41	31	26	16		
2012-10-06	8	10	7	1	4	3	12	17	58	59	...	114	96	76	73	55	38	18	15	20	19		
2012-10-07	6	12	2	4	1	6	9	14	43	67	...	115	109	93	73	45	23	36	35	9	11		

Données constituées de :

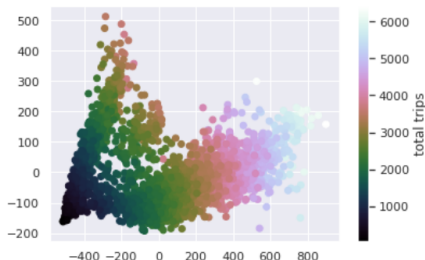
- 2920 jours
- 48 variables

Analyse en composantes principales

Utilisation de l'ACP

On procède à une ACP sur les données afin d'effectuer une réduction importante de dimension et de ne conserver que les dimensions portant le plus d'information. Ici, on souhaite que les composantes préservent 90% de leur variance.

Figure: Graphique obtenu après ACP

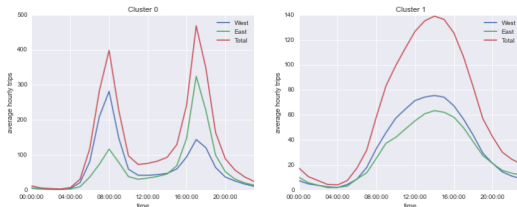


Clustering non supervisé

Utilisation du clustering

On utilise le clustering afin de séparer nos données en groupes distincts. Au vu de la distribution de nos données, nous avons décidé d'utiliser les mélanges gaussiens, qui semblent plus adaptés. On observe donc deux types de cluster.

Figure: Clusters



Des vendredis pas comme les autres?

Remarque

Certain vendredis sortent de notre classification, et semble n'appartenir à aucun groupe. Nous allons tenter de d'expliquer ce phénomène.

Figure: Représentation graphique des vendredis

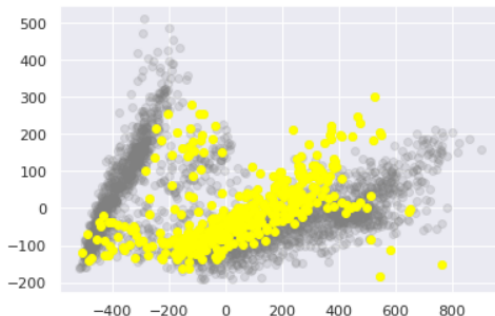


Table des matières

- 1 Les données
- 2 Interprétation des données
- 3 Conclusion**

Conclusion

Les cyclistes de Seattle ont tendance a :

- poser un jour de congé pour les fêtes nationales
- aller travailler les jours de fête nationale moins commune
- éviter de travailler les week-ends

[1] Joseph Salmon, *Modèle linéaire avancé : ANOVA*, 2019, <http://josephsalmon.eu/enseignement/Montpellier/HMMA307/ANOVA.pdf> ;

[2] *Learning Seattle's Work Habits from Bicycle Counts*, 2015, <https://jakevdp.github.io/blog/2015/07/23/learning-seattles-work-habits-from-bicycle-counts/>