Étude de cas 1 : Les polluants dans l'air intérieur

HASSINI Houda

Janvier 2019

1 Résultats ultérieurs

Le jalon 1 de l'étude de cas état un jalon exploratoire qui nous a permis de se familiariser avec les variables et détecter quelques pistes de problématique à explorer.
Lors de ce jalon, nous nous sommes basé sur une étude d'experts qui liait la présence de polluant tel que le formaldéhyde et le benzène à la présence de certaines colles (de types colles à bois, colle pour recouvrement de sols ...) et à l'usage de cigarettes. Nous nous sommes intéresse par la suite aux types de logement car les caractéristiques qui décrivaient ceux-ci était celle considérées par les experts comme sources des polluants.
L'étude de ces caractéristiques nous a permis dans un premier temps de voir qu'il y avait des caractéristiques qui ressortaient particulièrement chez certains individus plus que les autres. Par ailleurs, cette étude préliminaire nous a permis de voir que les individus étudiés était distribuées en 3 groupes par la variable année de construction.
Tous les indicateurs, nous permettaient de voir que nos données sont représentatives de plusieurs groupe à caractéristique différentes. Une classification non supervisée avec une CAH dans un premier temps qui nous a suggéré un découpage en 3 clusters, ensuite on a renforcé cette classification par une classification avec l'algorithme k-means. Le résultat obtenue est représenté sur les 2 premiers plans de AFDM dans la figure suivante:

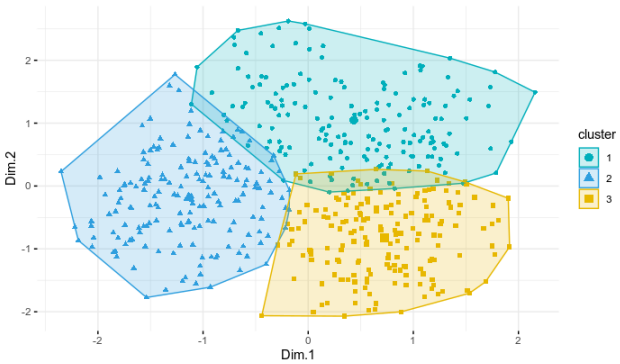

Figure 1: Résultats K-moyennes

Les clusters retrouvés sont caractérisé de la manière suivantes:

- **Groupe 1** : groupe des logements les plus anciens, caractérisé par des sols principalement sans revêtements ou couverts de parquets stratifié et de bois massif , des taux plafonds en tissu ou sans revêtements élevé, des taux de mus en peinture et sans revêtement élevé, un taux faible de murs en carrelage et un nombre d'appareils à combustions raccordées à un conduit fermé du logement très bas.
- **Groupe 2** : ce groupe ce caractérise par des logements un peu plus récents que ceux du groupe 1 mais très proche de la moyenne du l'ensemble. Il a des taux de sols en carrelage et en bois massif très bas contrairement au taux de sols en plastique. les taux de plafond sans revêtements, en bois et en plastique sont très faibles dans ce groupe et il est de même pour le taux des murs sans revêtements, en lambris en tissus et en crépie. Ces logement se caractérise par une élévation du taux des menuiseries en PVC et des meubles en meubles en bois massifs. Ils sont également équipés de très peu d'appareils à combustions raccordées à un conduit fermé du logement.
- **Groupe 3** : Il s'agit du groupe des logements les plus récents, les taux des sols sans revêtements et en plastique sont faible pour ces logements contrairement aux taux de sols en parquet stratifié et en carrelage, les plafonds en bois et en tissus sont faiblement présents , les revêtements du plafond les plus observés dans ces logements sont plafond en peinture. On observe aussi une dominance des meubles en bois massif et de menuiserie autres que le PVC et le bois, le nombre d'appareils à combustions raccordées à un conduit fermé du logement très élevé.

Quant aux variables quantitatives nous nous sommes intéressé de près au valeurs test. Les modalités significatives sont résumé dans le tableau ci-dessus, les modalités avec la valeurs test (v.test) positives sont sur-représentées dans le groupes, tandis ce que les modalités avec la valeurs test négatives sont sous-représentées. Par exemple le groupe 1 se distingue par une sur représentation des logements sans aération, le groupe 2 par une sur-représentation des logements équipés d'une cheminée et le troisième groupe se distingue particulièrement par une sur-représentation des logements à garage attenant.

Groupe 1

Modalités	v.test
MATER056=murs_princ_pierre	14,074363
MATER03=no_murs_princ_beton	9,147371
KVNT2e14=no_aeration	8,671110
HPLBO=plancher_pb_bois	7,733444
REAB23=renov+5ans	6,299647
DGG2be1=garage_attenant	6,286315
CHEM1=avec_chem	4,138963
DBRI=lieu_brico_coller	3,391679
HCU11=cuisine_fermee	-2,605936

Groupe 2

Modalités	v.test
CHEM1=avec_chem	14,058751
DBRI=sans_lieu_brico_coller	12,142468
HCU323=hotte_rejet_extr	12,041599
MATER056=murs_princ_pierre	6,931800
DCA3e1=cave_non_comm	6,742615
DGG2be1=sans_garage	6,251494
KVNT2e14=no_aeration	4,534630
REAB23=renov+5ans	3,292307

Groupe 3

Modalités	v.test
DGG2be1=garage_attenant	14,681198
CHEM1=sans_chem	8,292205
DCA3e1=cave_comm	5,111576
KVNT2e14=no_aeration	4,700703
Fumeurs.FUMEURn=aucun_fumeurs	3,251554
KCC1be1234=chauff_gaz	3,206747
REAB23=renov+5ans	3,183699
KCC1be1089=chauff_elec/autres	-3,933208
HPLBO=plancher_pb_beton/autres	-6,819811
HCU34=avec_hotte	-10,540358
DBRI=sans_lieu_brico_coller	-11,343736

Figure 2: Les modalités significatives pour chaque groupe

Nous avons aussi observés que la distribution de chaque polluants est différente d'un groupe à l'autre.

2 Objectif et problème

Dans ce jalon nous prévoyons de prédire le formaldéhyde en faisant ressortir aux mieux les caractéristiques de chacun des clusters.
La problème de base rencontrée lors de ce jalon est le fait que nous avons très peu d'observations par rapports au nombres de variables que nous avons eux grâce à l'AFDM. Pour y remédier nous avons choisi de faire une sélection de variable (un choix qui s'imposait à cause du nombre de variable : 62 variables pour seulement 170 observations en moyennes par cluster.

3 Sélection de variable

Nous avons fait une sélection de variable par la méthode **Stepwise** qui sélectionne les variables en supprimant des prédicteurs du modèle existant ou en en y ajoutant sur la base des résultats du test F. La méthode **Stepwise** combine les procédures de sélection ascendante et d'élimination descendante.
Nous avons choisi d'utiliser le critère **AIC** qui permet de pénaliser les modèles en fonction du nombre de paramètres afin de satisfaire le critère de parcimonie.nous avons fait le choix de se critère car l'AIC est asymptotiquement optimal lorsque l'on souhaite sélectionner le modèle avec l'erreur quadratique moyenne.

3.1 Sélection de variable du 1èr Cluster

L'application de l'algorithme précèdent permet de retenir les variables représentées dans le graphique suivant:

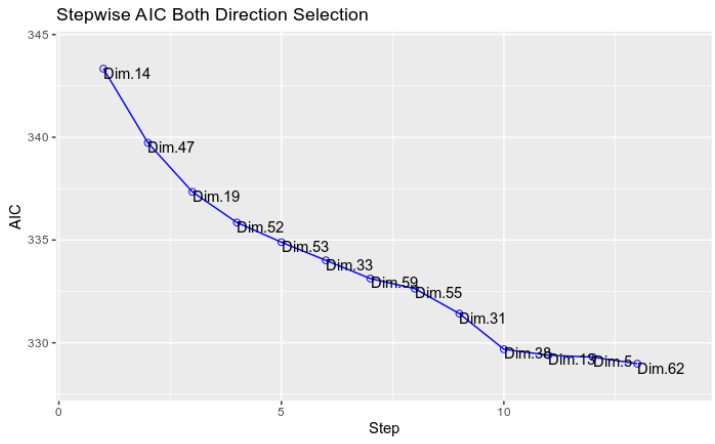

Figure 3: Variable sélectionné du 1èr Cluster

3.2 Sélection de variable du 2ème Cluster

L'application de l'algorithme précèdent permet de retenir les variables représentées dans le graphique suivant:

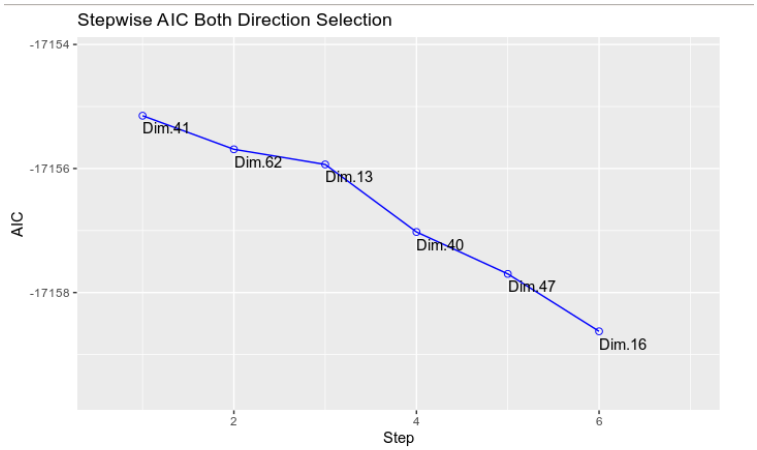

Figure 4: Variable sélectionné du 2ème Cluster

Sélection de variable du 3ème Cluster L'application de l'algorithme précèdent permet de retenir les variables représentées dans le graphique suivant:

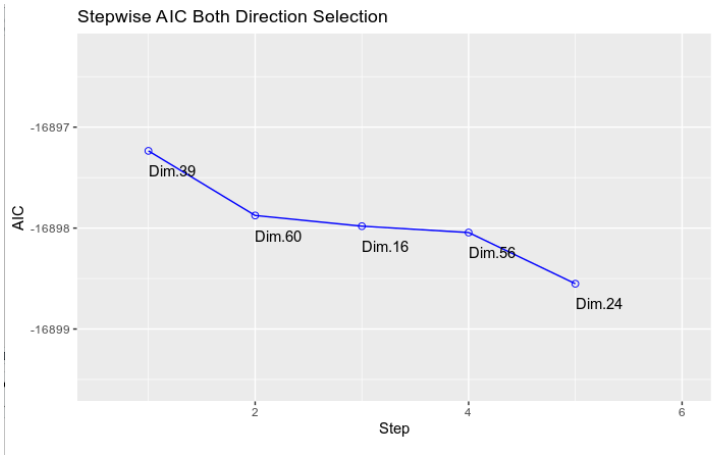

Figure 5: Variable sélectionné du 3ème Cluster

On peut voir le nombre de composantes sélectionnées varie d'un cluster résultat au quel on pouvait s'attendre étant donner que notre variable cible ne se distribue pas de la même façon pour chaque cluster.

4 Prédictions

Pour prédire le Formaldéhyde nous utilisons deux modèles : Un K plus proches voisins (KNN) et un SVM radiale. Le choix de ces deux algorithme est dû au nombre très faible d'observations par rapport à celui des variables nous obtenons les résultats pour chacun des cluster est le suivant: On peut voir que les résultats obtenus par les deux algorithmes

	train	test
KNN	0.84	0.90
SVM	1.12	1.15

Figure 6: RMSE en train et en train et en test pour le premier cluster

	train	test
KNN	0.87	1.02
SVM	0.85	1.05

Figure 7: RMSE en train et en train et en test pour le deuxième cluster

	train	test
KNN	0.86	0.94
SVM	0.86	0.89

Figure 8: RMSE en train et en train et en test pour le troisième cluster

est prometteurs en termes de performances. Cependant es résultats changent d'un classe à l'autre.

5 Conculision et Jalon 3

Les résultats de ctte parties sont des résultats prometteurs mais ils est indispensables de les analyser ce qui fera partie du jalon 3. Les méthodes utilisées ainsi que les algorithmes peuvent potentiellement améliorer :

- Une meilleur sélection de variables par arbres de décision par exemple afin de prendre en compte les liaisons non linéaires lors de la sélection de variable.
- Tester d'autres algorithme de sélection.
- Expliquer les résultats obtenues en fonction des variables initailes et faire le liens avec les études des experts