1. ANALYSE STATISTIQUE DES DONNÉES.

A_ Analyse globale.

Nous traitons dans cette première partie l'analyse statistique des données. Puisque certains points possèdent des données manquantes, nous avons décidé de simplement les retirer du jeu de données.

Une première étude de la variance de chacune des variables (figure 1) nous permet de repérer certaines valeurs extrêmes, que nous décidons de retirer également du jeu de données. Nous obtenons donc un boxplot ajusté que nous pouvons étudier plus correctement (figure 2).

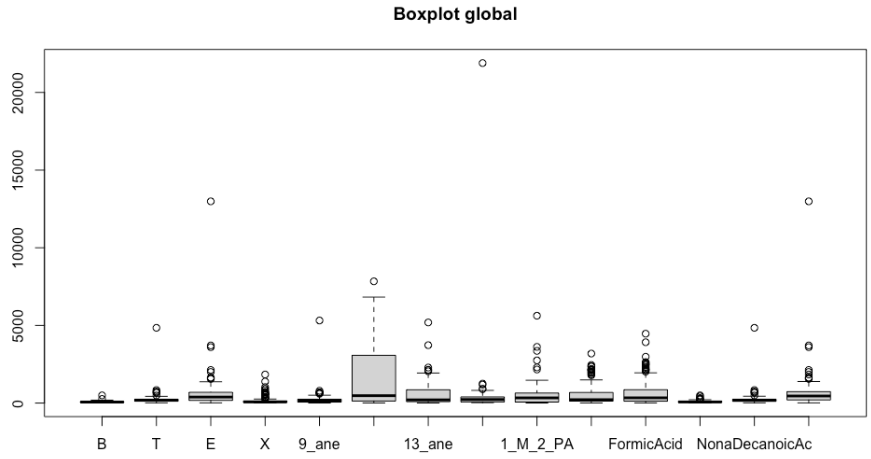


Figure 1

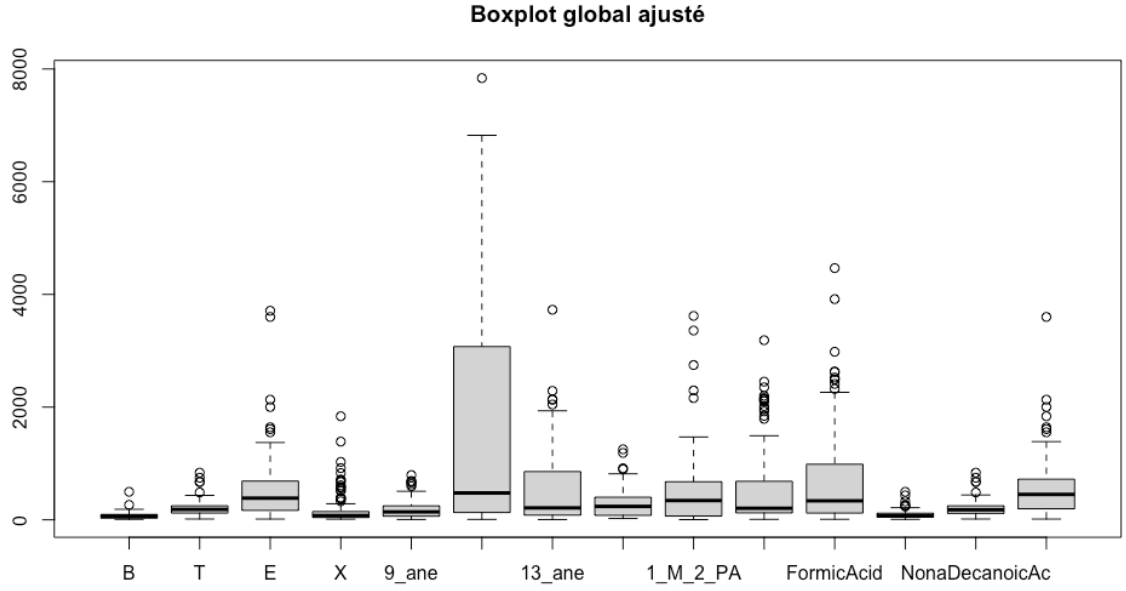


Figure 2

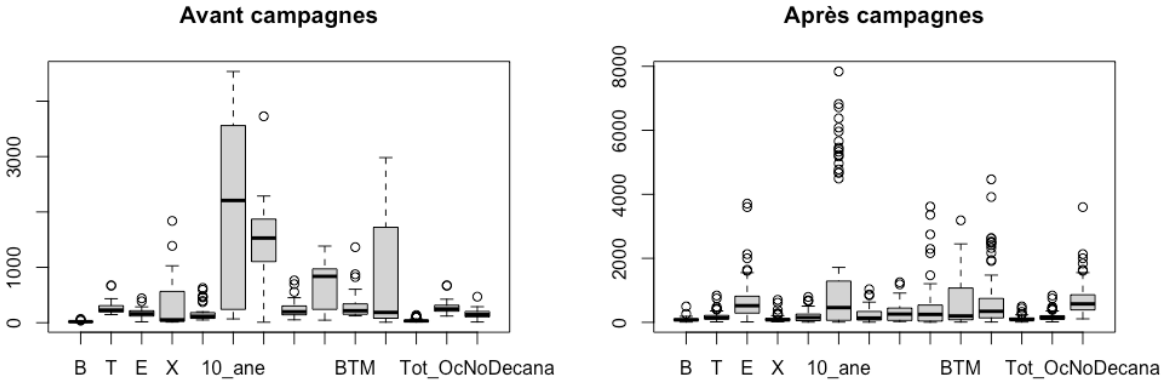
Puisque les données ne sont pas normées, on remarque rapidement que quelques composés possèdent sur l'ensemble des campagnes une variance prédominante par rapport aux autres, c'est le cas par exemple du 10_ane. Les médianes de chaque composé semblent quant à elles toutes du même ordre de grandeur, bien que les unités ne soient évidemment pas les mêmes.

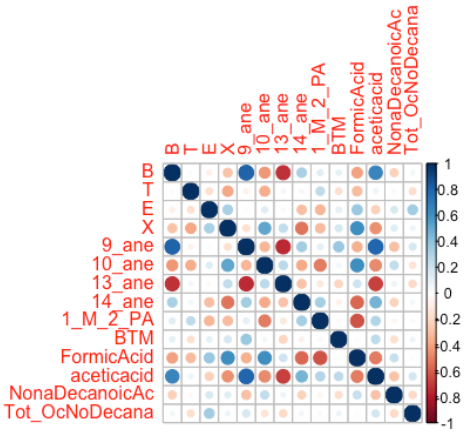
B_ Analyse par campagne.

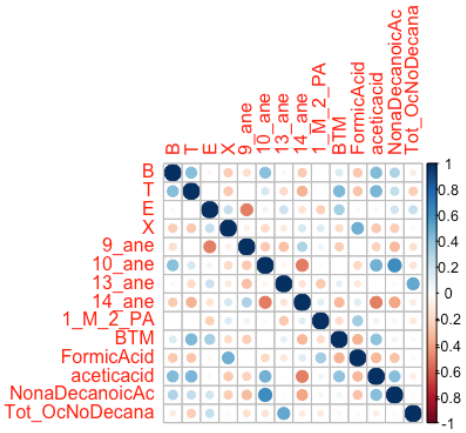
L'analyse par campagne nous permet d'identifier une présence plus forte de certains composés, notamment le 10_ane et le 13_ane, qui apparaissent en grosse proportion sur l'ensemble du territoire testé avant l'implantation du site (campagnes BF1 et BF2) mais qui semblent moins présents après l'implantation (campagnes CA1 à CA4). On compte cependant plus de mesures extrêmes de 10_ane après l'implantation.

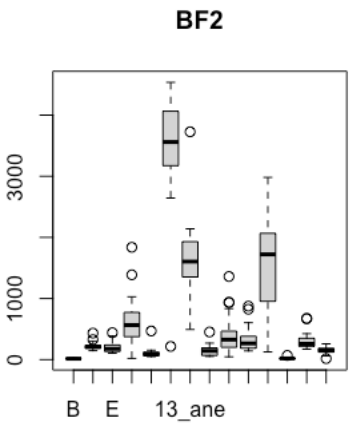
Les composés sont également plus corrélés avant l'implantation du site. On constate effectivement que le corrélogramme pâli globalement pour l'ensemble campagnes après implantation. La présence de benzène par exemple, très inversement corrélée à celle de 10_ane sur les campagnes BF est n'est quasi plus corrélée sur les campagnes CA.

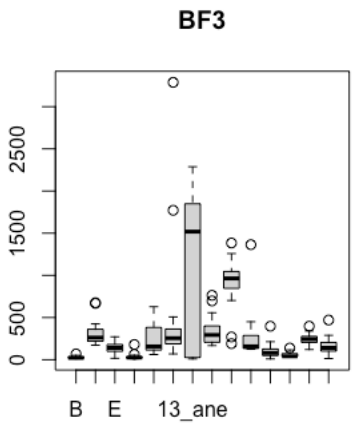
On remarque cela dit quelques similitudes entre les campagnes réalisées avant et celles après, c'est le cas par exemple des campagnes BF1 et CA2 pour lesquelles l'acide formique semble très présent sur l'ensemble des points de prélèvement. Les deux campagnes ayant été réalisées en été, il est intéressant de s'intéresser à l'analyse statistique des variables suivant la discrimination été/hiver.

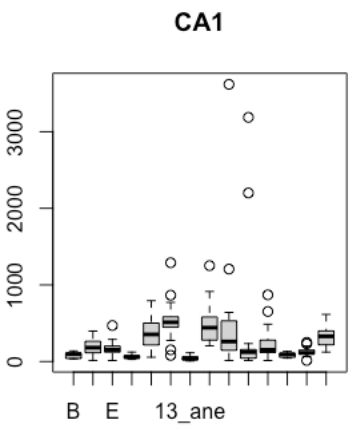


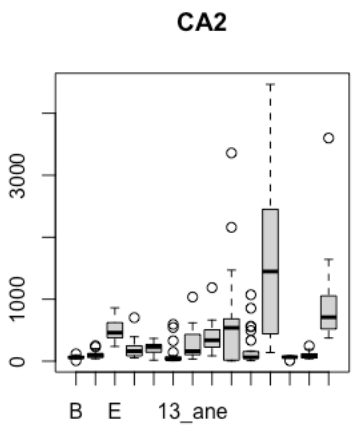


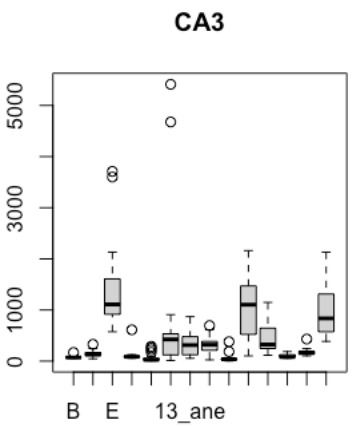


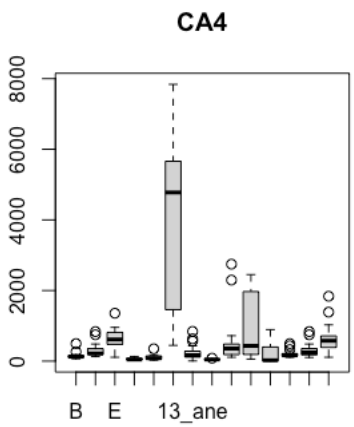

BF2
0
1000
3000
B
E
13_ane


BF3
0
500
1500
2500
B
E
13_ane


CA1
0
1000
2000
3000
B
E
13_ane

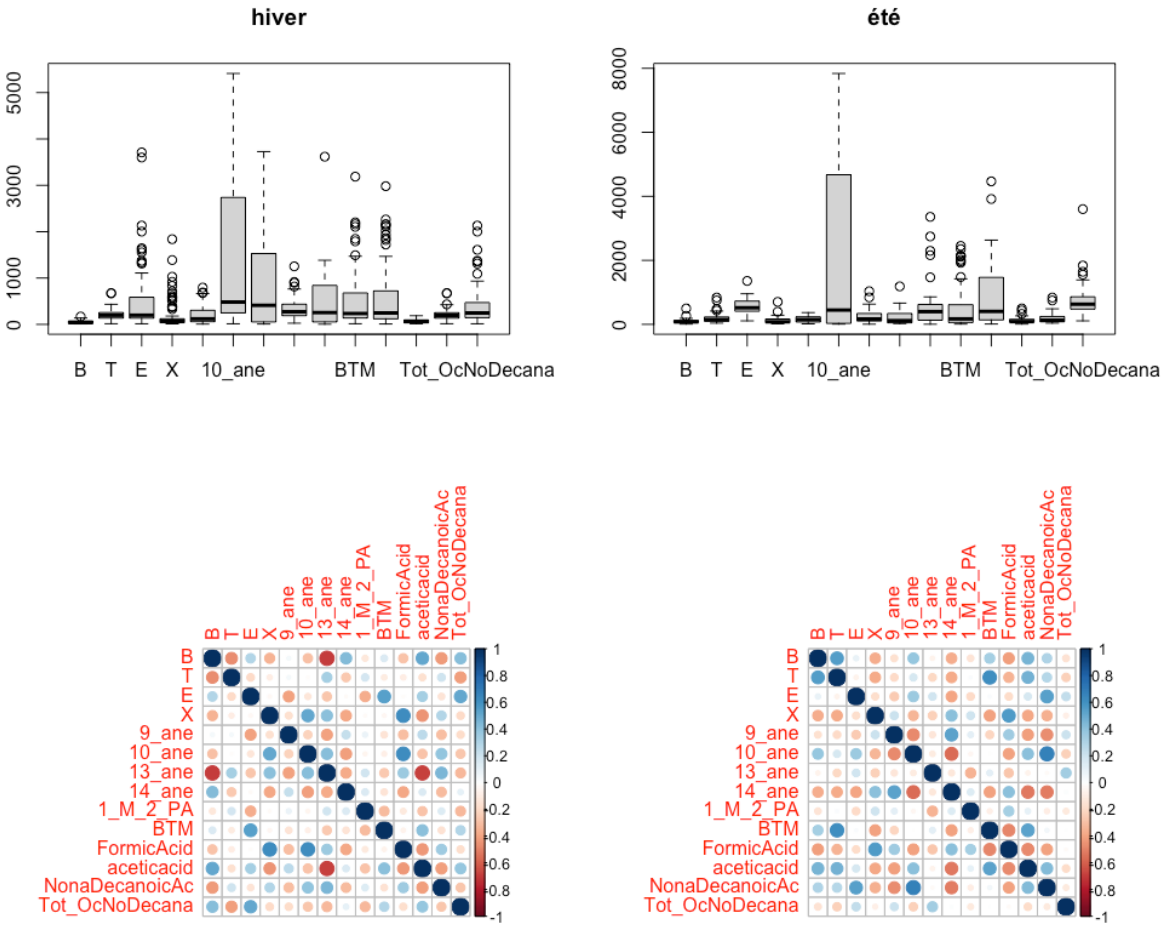

CA2
0
1000
3000
B
E
13_ane


CA3
0
1000
3000
5000
B
E
13_ane


CA4
0
2000
4000
6000
8000
B
E
13_ane

C_ Analyse par saison.

L'analyse des variables par saison nous indique que la saison possède une forte importance sur la proportion de 13_ane présent par exemple alors que le 10_ane lui n'est quasiment pas impacté. On observe également globalement moins de valeurs extrêmes en été qu'en hiver, phénomène particulièrement visible pour l'éthylbenzène par exemple.



D_ Commentaire général.

L'analyse statistique des variables nous permet d'établir plusieurs conclusions : on remarque que certains composés, comme le 13_ane, sont fortement impactés par la saison de l'année, et de la même manière que l'installation de l'usine a impacté leur présence. Au global, on observe davantage de valeurs extrêmes de 10_ane et de Tot une fois l'usine implantée, ce qui pourrait signaler une production de ces composés par l'usine.

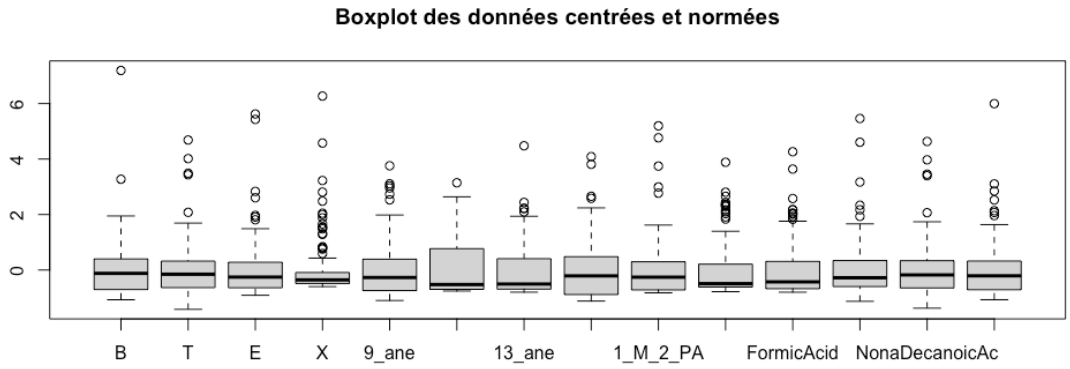
De par l'importante variance de quelques composés, il sera nécessaire de normer nos données pour éviter d'avoir un nuage de points trop étirés dans certaines directions.

2. ACP ET RÉDUCTION DE DIMENSION

A_ Analyse

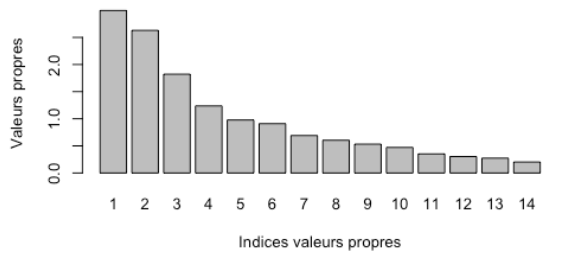
Comme analysé précédemment, il est important de centrer et de réduire nos données afin d'éviter que certains composés de très forte variance prennent le dessus sur les premières composantes de notre ACP.

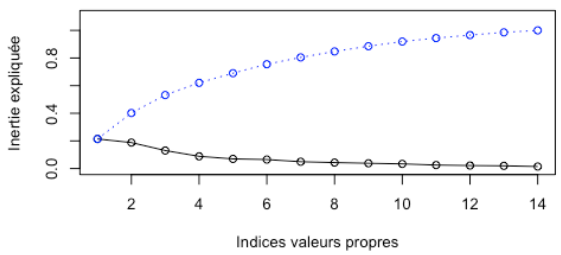
On peut alors tracer le boxplot des données centrées et normées ci-dessous.

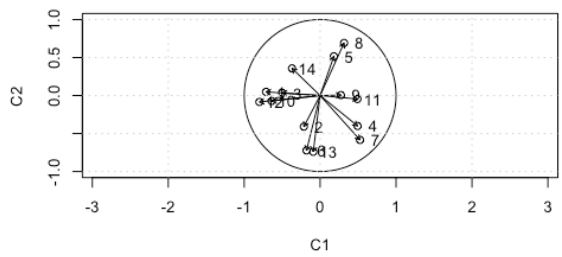


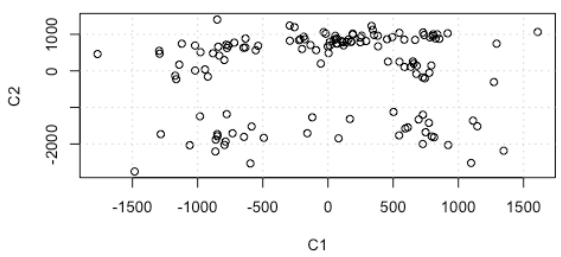
La question ici est de savoir s'il est possible de réaliser une réduction de dimension convenable. L'analyse de l'ACP normée produit les graphiques ci-dessous. Remarquons plusieurs éléments :

- Aucune valeur propre ne semble réellement dominer les autres en termes de pourcentage d'inertie expliquée. La question du choix du nombre d'axes se pose alors : respecter le critère de Kaiser ? Choisir un nombre d'axe permettant d'expliquer une majeure partie de l'inertie totale ?
- Les composantes du premier plan factoriel expliquent fortement certaines variables, et plusieurs d'entre elles paraissent indépendantes. Il serait donc possible de donner une signification à ces deux axes.

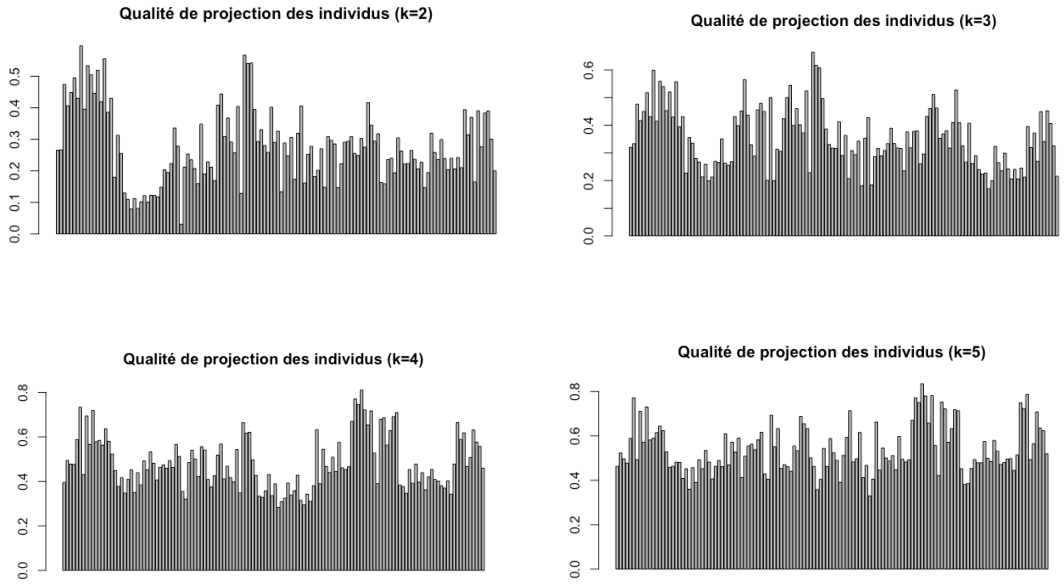




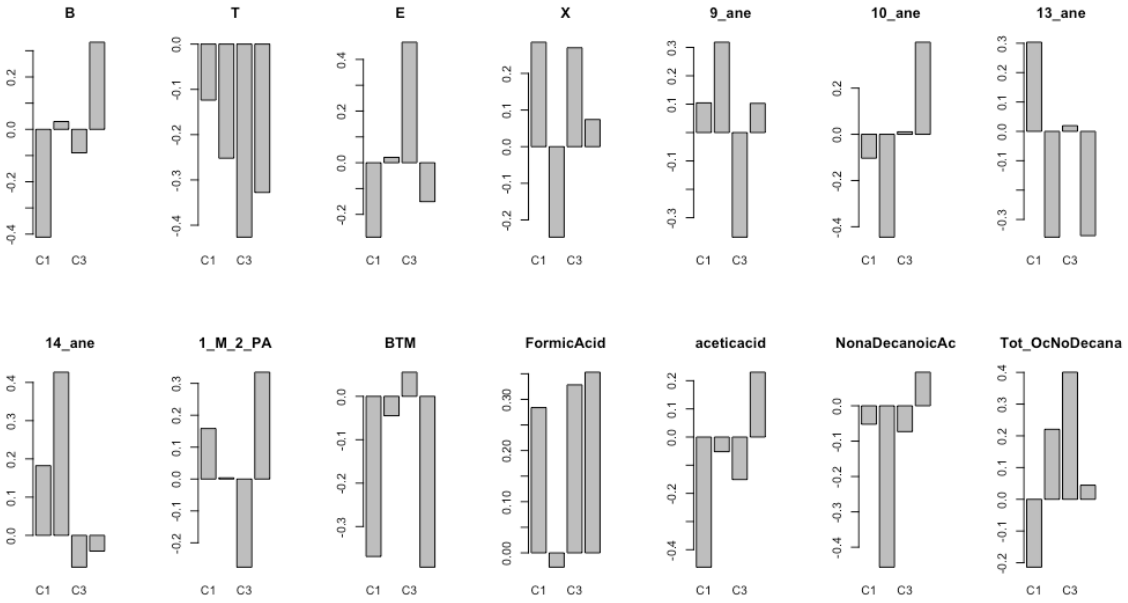




Nous décidons de tracer la qualité de chacun des individus en fonction du nombre de composantes retenues pour l'ACP. Le graphique ci-dessous nous permet alors de rejeter les dimensions k=2 et k=3, car les individus ne sont pas suffisamment bien projetés. On remarque également qu'au-delà de la dimension k=4, l'amélioration de la qualité de projection ne se remarque pas nettement. On choisit donc de travailler dans un espace réduit avec les 4 premières composantes principales, ce qui revient d'ailleurs à suivre la règle de Kaiser sur les valeurs propres. En suivant cette méthode, on n'arrive à capter cependant qu'environ 60% de l'inertie totale du nuage.



On lit sur le cercle des corrélations que certains composés sont particulièrement bien expliqués par le premier axe factoriel : c'est le cas de l'acide formique, positivement corrélé à l'axe C1, et du benzène, de l'acide acétique et du BTM, tous négativement corrélés à l'axe C1. Le diagramme ci-dessous trace la contribution de chacune des variables aux 4 premiers axes factoriels.



On remarque que le 10_ane est négativement corrélé à l'axe C2 et positivement à l'axe C4, de la même manière que la variable appelée « Tot ». On pourrait donc probablement associer ces axes à nos deux campagnes de mesure BF et CA : le 10_ane et le Tot sont effectivement deux des composés les plus impactés par l'implantation de l'usine.
De la même manière les composés X, le 13_ane et l'acide acétique sont trois composés fortement impactés par la saison, l'axe C1 pourrait donc être associé à la saison de l'année.

Représenter les plans factoriels engendrés par (C1, C2) puis (C2, C4) nous permet d'ailleurs de vérifier cette interprétation :

- On remarque effectivement une séparation plus ou moins marquée entre les points mesurés en hiver et ceux en été sur l'axe C1 (figure 3).
- La séparation est plus marquée encore concernant les mesures réalisées avant l'implantation de l'usine puis après, puisque l'on distingue clairement 2 groupes de points (figure 4).

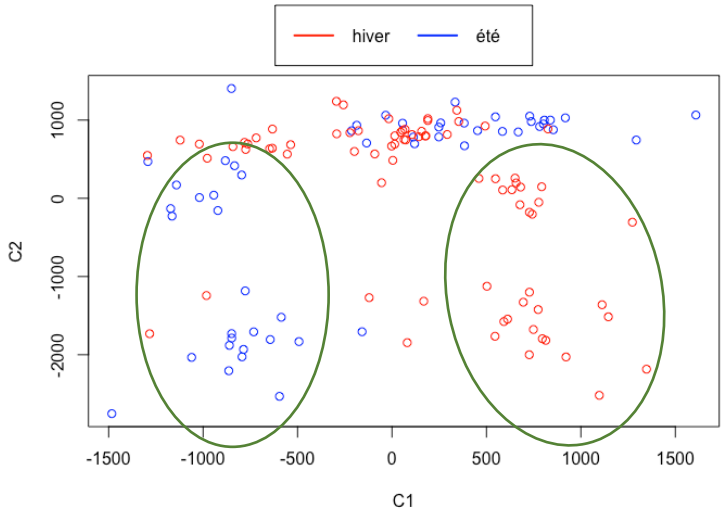


Figure 3

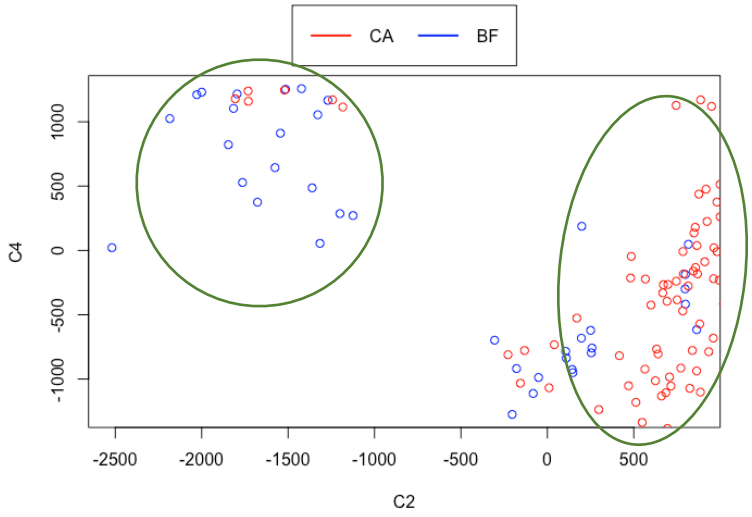


Figure 4

B_ Conclusion

L'analyse en composantes principales nous permet de faire ressortir 4 composantes qui expliquent relativement bien l'ensemble des individus et qui permettent de les classifier selon la saison (été ou hiver) et la temporalité des mesures (avant ou après installation). On peut donc effectivement faire ressortir une signature des composants pour chaque période.

Cela dit, bien que l'ACP puisse être une première méthode permettant de mettre en évidence ces signatures, les résultats que l'on obtient ne sont pas pour autant optimaux. En effet le nombre de dimensions choisies reste très arbitraire et les distinctions entre les différents individus ne sont pas extrêmement nettes, bien que visibles. Nous ne nous sommes pas intéressés aux points de mesure ni aux conditions météorologiques lors de ces mesures, qui peuvent avoir un impact sur notre analyse et que nos composantes ne font pas forcément ressortir.

Notons tout de même que les concentrations de 10_ane, 13_ane et de xylène sont des variables clés dans l'identification de la signature des individus pour chaque période.