

Rapport

Analyse des données participatives de la pollution de l'air en lien avec la santé via un questionnaire

A titre du projet TER du Master 1 informatique de L'université de
Versailles-Saint-Quentin-en-Yvelines

Dans le cadre du projet de POLLUSCOPE :

Observatoire participatif pour la surveillance de l'exposition
individuelle à la pollution de l'air en lien avec la santé

Encadré Par :

- Mme Karine Zeitouni
- Mme Hafsa ELHAFYANI

Réalisé par :

- M. EL ASSRI Aziz
- M. LI Arnold

Résumé

En tant que des informaticiens, nous sommes censés à assurer un system d'information fiable qui permet de prendre des décisions pertinentes qui affectent à grande échelle la population, c'est le fait qui exige de faire des analyses descriptives et statistiques.

A partir de ce point, et en utilisant les outils d'analyse informatiques, on s'est engagé d'étudier les données réelles collectés par des micro-capteurs mis à disposition des individus volontiers et qui mesurent l'exposition à la pollution atmosphérique pendant les activités journalières de ces participants,

Sommaire

Résumé.....	2
Sommaire.....	3
1. Introduction.....	4
2. Objectif global	5
3. Principales études réalisées.....	5
3.1. Les polluants en fonction des catégories	5
3.2. Etude polluant en fonction des participants	6
3.3. Etude sur la fiabilité.....	6
3.4. Les polluants en fonction des activités	7
3.5. Etude sur la corrélation.....	7
3.6. Les polluants en fonction des événements	7
3.7. Analyse de Variance (anova)	8
3.8. Outliers.....	8
3.9. Mini-application.....	9
4. Outils utilisés	10
5. Problèmes rencontrés.....	10
6. Perspectives.....	11
7. Conclusion	11
Bibliographie	12
Annexes.....	13

1. Introduction

Dans la vie de tous les jours, nous sommes tous exposés plus ou moins à de la pollution que ce soit chez soi ou même au travail mais cependant chaque personne est exposé différemment au même polluant, c'est pourquoi en utilisant un questionnaire auquel plusieurs participant ont répondu, on va essayer de comprendre pourquoi.

Dans cette étude on va étudier les PM1.0, les PM2.5, les PM10, le BC et le NO2. Les PM sont des particules fines et ultrafines dans l'air qui peuvent à un taux élevé avoir un risque sanitaire sur la population, le black carbone(BC) est aussi un polluant de l'air qui est essentiellement émis par les pots d'échappement et par la combustion domestique et est également dangereux pour la santé puis le dioxyde d'azote(NO2) est un polluant majeur de l'atmosphère terrestre produit par les moteurs à combustion interne et les centrales thermiques.

2. Objectif global

Analyser les données participatives d'un questionnaire en relation avec les mesures des polluants élaborés par des capteurs mis à disposition aux participants volontiers.

Pour mener notre analyse à notre objectif global, on s'est fixé plusieurs sous objectifs:

- Pouvoir comprendre des données à partir d'un questionnaire et de données fournies.
- Faire le lien entre les données de mesure et de questionnaire.
- Pouvoir les analyser et savoir interpréter les résultats obtenus.
- Visualiser les résultats obtenus sous forme des diagrammes.

3. Principales études réalisées

3.1 Les polluants en fonction des catégories

On a d'abord fait une étude générale c'est-à-dire que l'on a fait des requêtes visant tous les participants du questionnaire en fonction de leurs catégories comme on peut le voir sur le premier rapport rédigé. (Annexe 1 page 13)

On a fait des statistiques générales (min, max, moyenne, médiane, écart-type...etc) sur tous les polluants pour chaque catégorie. On remarque que les non sportif et les personnes 35 à 50 ans sont les plus exposés aux particules fines, et la même chose pour les non-fumeurs concernant le polluant BC.

3.2 Les polluants en fonction des participants

Ensuite dans le rapport2, on a calculé les moyennes des polluants en fonction des participants et on avait remarqué que c'est le participant 9999932 qui était le plus exposé mais qu'il avait aussi des valeurs aberrantes et qu'il était aussi le seul non sportif du coup on s'est aussi intéressés aux participantx 9999944 et 9999946, car ils avaient une forte exposition avec pratiquement aucune valeur aberrante (Annexe2 page 14).

3.3 Etude sur la fiabilité

On s'est posé la question sur la fiabilité des données des mesures d'une part, et d'autres part pour la logique des correspondances entre les différentes variables, et en plus on s'est demandé la fait d'avoir les questionnaires de seulement quelques participants qu'on a va poser des problèmes concernant la fiabilité de nos analyses, du coup on a décidé d'étudier les participant qu'on a leurs questionnaires en faisant la jointure entre les tables avant chaque étude.

Comme on a trouvé un grand nombre d'anomalies surtout concernant la correspondance entre l'activité et l'événement dont les mesures ont été prise par le participant. Pour traiter ces anomalies, on les a listé un fichier 'csv ' les correspondances logiques entre l'activité et événement, puis on a utilisé ce fichier pour appliquer des requêtes et enlever ces anomalies de la table mesures (Annexe 3 page 15).

3.4 Les polluants en fonction des activités

Dans le rapport 3, on a détaillé comment on a calculé les moyennes des polluants en fonction des activités, on a remarqué que dans les restaurants, il y avait une forte exposition aux polluants (Annexe 4 page 16).

3.5 Etude sur la corrélation

On a essayé de chercher s'il existait des corrélations entre les différentes données et en effet entre les différents particules ce qui est évident puisqu'ils ne se différencient pas par leur taille et on a trouvé que les autres corrélations étaient soit trop faibles ou très peu intéressantes (exemple : entre latitude et humidité). Annexes 5,6 page 17

3.6 Les polluants en fonction des événements

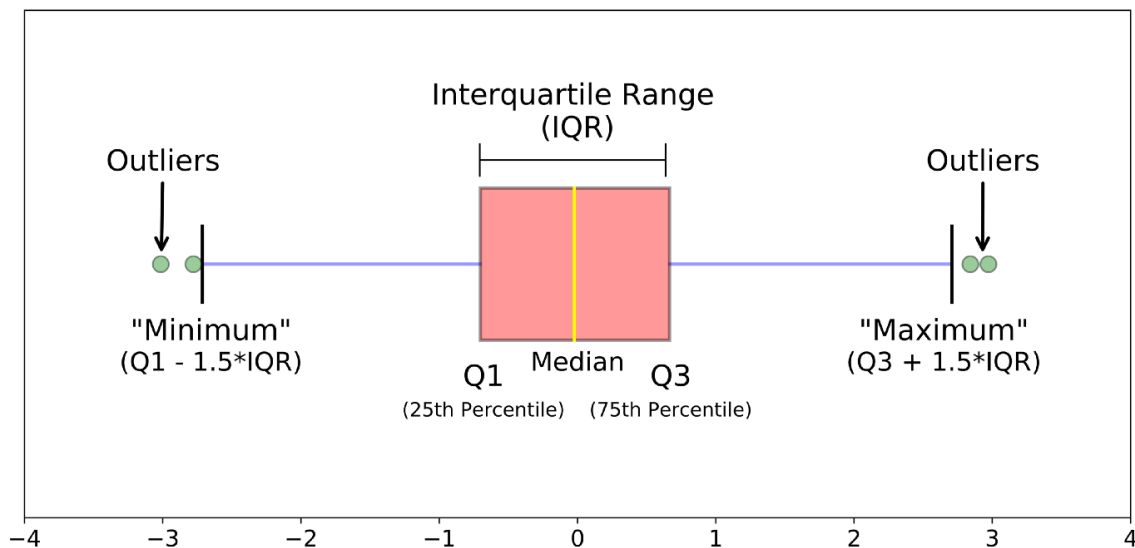
D'après les plots sur le pourcentage de chaque polluant par rapport à l'événement, on en a déduit que les événements allumages de cheminée et marcher étaient ceux qui produisent le plus de polluant. Ce qui paraît logique puisque marcher on est forcément dehors, ce qui rend le participant plus exposé au polluant de l'air, et allumage de cheminée qui crée de la fumée qui est un signe que le bois brûle mal et donc qui génère des particules cancérogènes.

(Annexes 7,8 page 18)

3.7 Analyse de Variance (anova)

On a appliqué un modèle d'Anova à deux facteurs pour expliquer les variables des polluants avec des variables qualitatives; événement et activité ; on a trouvé que tous les polluants dépendent fortement de l'activité et l'évènement sauf le BC qui a une valeur de >0.05 selon le test de Fisher. (Annexes 9,10 page 19)

3.8 Outliers



Une valeur aberrante ou outlier est une valeur extrême de la distribution d'une variable, On définit les valeurs extrêmes comme les valeurs supérieures ou inférieures à 1.5 fois (généralement) l'écart interquartile. Ces valeurs peuvent affecter la fiabilité des analyses et se produisent quand un utilisateur ne répond pas à une ou plusieurs questions ou à cause du mal saisie, et pour pallier à cela nous avons utilisé la méthode standard de R qui détermine ces valeurs par rapport aux quartiles (25% et 75%).

Les boxplot permet d'identifier les outliers, on a fait un traitement sur le polluant BC, vous les trouverez dans le dossier Plots/Outliers sur notre dépôt GitHub <https://github.com/LI-Arnold/TER>

Le traitement de ces valeurs consiste à les remplacer le médian ou bien par le Q3(75%) s'ils se trouvent au-dessus du max(1.5 écart interquartile) et s'ils se trouvent au-dessous du min on les remplace par le Q1 (25%), et cela dépend des variables étudiées car on peut préciser éventuellement un intervalle des valeurs qui nous paraît à nous même logiques et possible, et c'était bien le cas pour l'humidité et la température. (Annexes 11, 12 page 20)

3.9 Mini application

On a utilisé le markdown sous Rstudio avec Shiny pour faire une démonstration interactive qui permet à l'utilisateur de visualiser les moyens des polluants selon l'activité et avec des diagrammes, en plus il a la possibilité de chercher un participant pour savoir plus d'informations sur lui, ce qui utile pour donner des explications sur les mesures prises par un tel participant. (Annexes 13,14 page 21)

4. Outils utilisés

Les outils qu'on a utilisés sont:

- R pour le langage de programmation car le langage R est très pratique et puissant pour faire des statistiques et analyser les données.
- SQLite pour les traitements des requêtes, et les dataframes pour le stockage des tables.
- Shiny pour faire une mini application de notre projet pour avoir de l'interaction avec l'utilisateur.
- Rmarkdown pour la rédaction des rapports.

5. Problèmes rencontrés

Au début, nous avons remarqué qu'il y avait des anomalies sur les données fournies car il est difficile de se fier à des données incomplète ou même fausse.

Les valeurs manquantes NA génèrent des problèmes quand on fait les requêtes sql, et on a du a faire des conversions entre des types différents (factor et char/num et int) sur R pour faire nos requêtes, en effet, il faut les traiter séparément avec une démarche spécifique comme les outliers, alors ce n'était pas un sujet pour notre étude, donc on a utilisé l'option « not null ».

6. Perspectives

Ce qui pourrait être amélioré ou fait si on avait plus de temps :

- On pourrait améliorer la mini-application pour aborder toutes les parties qu'on a étudié et ajouter des fonctionnalités supplémentaires comme une option pour générer les rapports automatisés ou corriger les valeurs aberrantes directement sur l'application et télécharger les résultats sous un format de données csv ou excel.
- Aborder pratiquement les données de GPS, pour localiser les endroits qui sont susceptibles d'être dangereux pour la santé, et permettre à l'utilisateur de savoir s'il vit ou pas dans un environnement pollué.

7. Conclusion

Ce projet nous a beaucoup appris et a été très enrichissant que ce soit au niveau des connaissances ou le travail d'équipe. Il nous a permis aussi de découvrir les enjeux liés à la pollution dans notre planète, ses risques qui menacent la santé de l'humain. Grâce à ce projet, on aimerait que nos futurs projets soient encore mieux !

Bibliographie

[1] [page officielle du projet polluscope]. <http://polluscope.uvsq.fr/index.php/fr/>

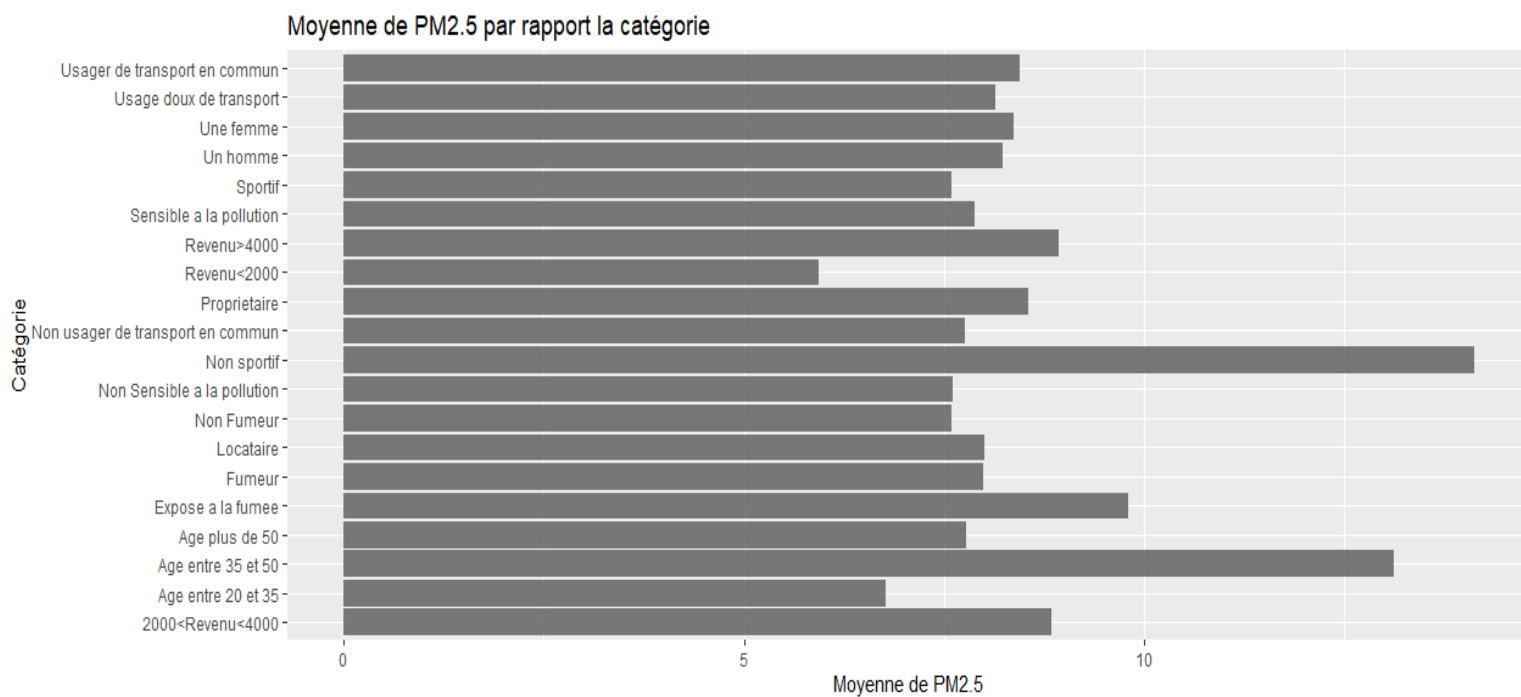
[2] Analysis of Variance (ANOVA) in R. Youtube :
<https://www.youtube.com/watch?v=qrP7evoNCy4>

[3] [les outliers]. Comment détecter les outliers avec R. <https://statistique-et-logiciel-r.com/comment-detecter-les-outliers-avec-r/>

[4] [Documentation markdown]. Create an attractive online dashboard using R,
<https://www.youtube.com/watch?v=H64zJqmzrMs>

[4] [Résolution des erreurs de codage]. <https://stackoverflow.com/>

Annexes







Annexe 1 moyenne du polluant PM2.5 par catégories

<div> <div>PolluantParCategorie x</div> <div>activity x</div> <div>age x</div> <div>p9999932 x</div> <div>moy_participant x</div> <div>Q moc</div> </div>						
<div> <div>Filter</div> <div>Q</div> </div>						
	Participant	MOY_PM2.5	MOY_PM10	MOY_PM1.0	MOY_NO2	MOY_BC
1	999994	4.584440	4.863378	3.352941	8.844402	NA
2	9999915	6.974502	7.848606	4.698805	7.100398	857.9785
3	9999916	8.811857	9.686467	5.622443	6.729342	723.9474
4	9999932	14.083665	18.877109	6.020436	13.752435	NA
5	9999941	5.701992	6.264570	3.657338	7.827532	877.9504
6	9999944	11.635490	14.201648	7.454951	7.718552	1001.6517
7	9999946	10.097119	12.104070	7.139533	NA	NA
8	9999964	7.009598	7.692412	5.140822	6.988929	NA
9	9999965	5.826164	6.375000	3.949190	12.830729	NA
10	9999966	5.999250	6.746784	3.544383	11.930853	NA
11	9999967	6.242507	7.018236	4.484140	NA	NA

Showing 1 to 11 of 11 entries, 6 total columns

Annexe 2 moyennes des polluants en fonction des participants qui ont répondu au questionnaire

	activity 	event 	nbr 
1	Bureau	Allumage De Chemin��e	791
2	Bureau	Arr��ter De Cuisiner	545
3	Bureau	Cuisiner	417
4	Bureau	Marcher	3155
5	Bus	Arr��ter De Cuisiner	585
6	Bus	Cuisiner	154
7	Bus	Fermeture De Fen��tre	218
8	Bus	Fumer	16
9	Bus	Marcher	87
10	Bus	Ouverture De Fen��tre	61
11	Cin��ma	Fermeture De Fen��tre	120
12	Domicile	Marcher	6952
13	Magasin	Fermeture De Fen��tre	62
14	Magasin	Ouverture De Fen��tre	20
15	Magasin	Sport	131
16	M��tro	Fermeture De Fen��tre	17
17	M��tro	Ouverture De Fen��tre	28
18	NULL	Fermeture De Fen��tre	143
19	NULL	Marcher	1
20	NULL	Ouverture De Fen��tre	1584

Showing 1 to 23 of 39 entries, 3 total columns

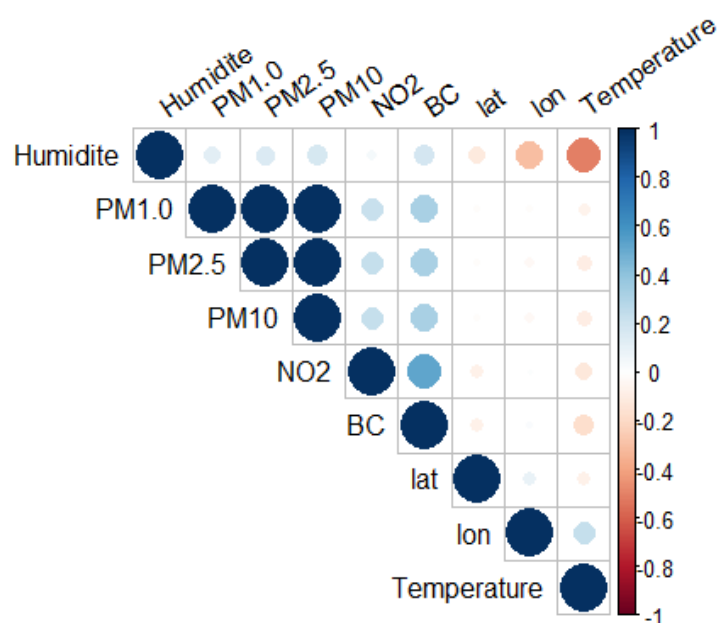
Annexe 3 :
Nombre d'anomalies
entre l'activit   et
l  v  nement

	participant_virtual_id	Domicile	Bureau	Rue	Voiture	Vélo	Bus	Train	Restaurant	Métro	Magasin	Parc
1	999994	NA	4.86	NA	NA	NA	NA	NA	NA	NA	NA	NA
2	9999915	13.73	NA	38.84	9.40	NA	NA	NA	NA	NA	NA	NA
3	9999916	8.26	6.72	5.74	9.25	NA	NA	NA	54.92	NA	NA	NA
4	9999932	4.12	19.84	118.41	9.97	NA	11.83	7.88	73.72	NA	2.60	NA
5	9999941	6.54	2.27	13.17	5.41	NA	NA	NA	64.47	NA	7.33	3.62
6	9999944	12.51	0.79	4.31	9.11	NA	4.20	NA	23.43	NA	7.09	NA
7	9999946	12.16	5.62	4.88	24.67	NA	NA	12.44	57.87	NA	3.31	NA
8	9999964	8.65	7.00	NA	11.58	NA	NA	NA	NA	NA	4.89	6.00
9	9999965	4.45	3.55	14.82	NA	NA	25.56	3.58	NA	NA	9.81	NA
10	9999966	5.38	NA	15.71	NA	NA	1.11	21.71	NA	NA	7.53	7.20
11	9999967	6.64	6.02	16.40	10.70	NA	26.93	3.74	NA	NA	13.27	NA

Annexe 4 : Moyennes des polluants en fonction des participants et selon l'activité

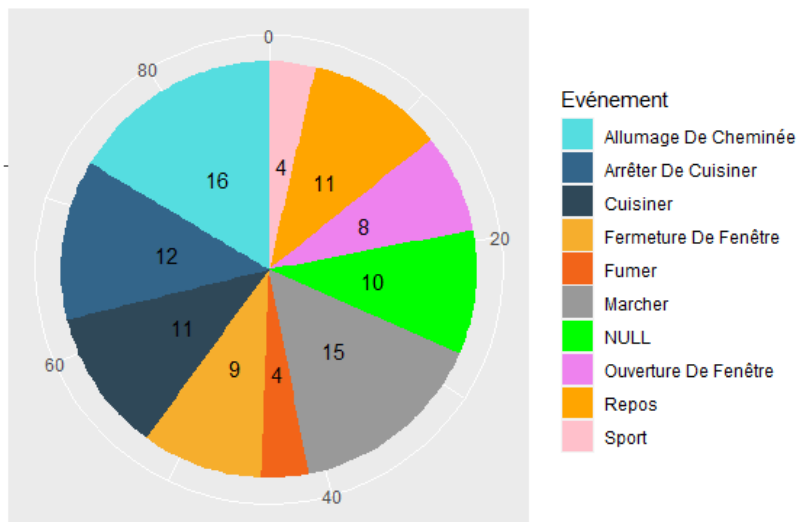
	lat	lon	PM2.5	PM10	PM1.0	Temperature	Humidite	NO2	BC
lat	1.00	0.09	-0.02	-0.02	-0.02	-0.08	-0.12	-0.08	-0.08
lon	0.09	1.00	-0.04	-0.04	-0.02	0.23	-0.31	0.01	0.02
PM2.5	-0.02	-0.04	1.00	1.00	0.99	-0.10	0.16	0.23	0.32
PM10	-0.02	-0.04	1.00	1.00	0.99	-0.10	0.17	0.23	0.32
PM1.0	-0.02	-0.02	0.99	0.99	1.00	-0.07	0.13	0.22	0.32
Temperature	-0.08	0.23	-0.10	-0.10	-0.07	1.00	-0.51	-0.13	-0.18
Humidite	-0.12	-0.31	0.16	0.17	0.13	-0.51	1.00	0.04	0.18
NO2	-0.08	0.01	0.23	0.23	0.22	-0.13	0.04	1.00	0.52
BC	-0.08	0.02	0.32	0.32	0.32	-0.18	0.18	0.52	1.00

Annexe 5 : Matrice de corrélation entre les différentes variables



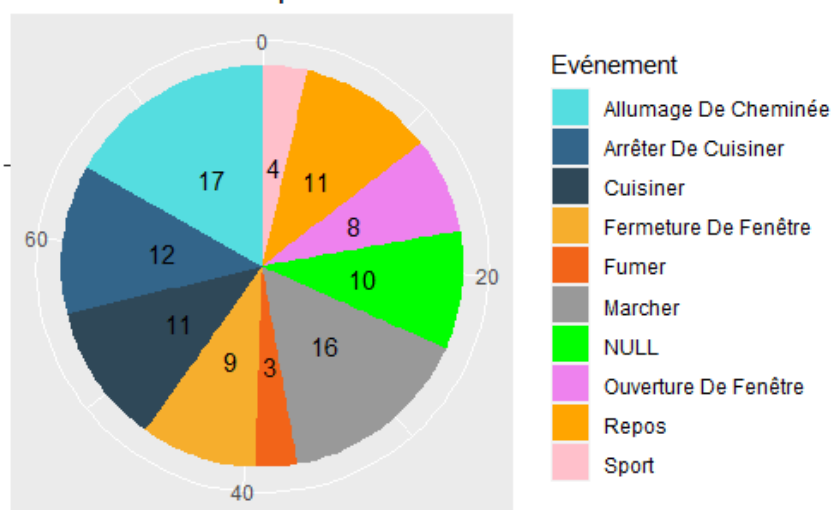
Annexe 6 : Plot de corrélation pour interpréter l'annexe 5

Mesures de PM10 par event en %



Annexe 7 : Produit PM10 selon chaque événement

Mesures de PM2.5 par event en %



Annexe 8 : Produit PM10 selon chaque événement

```

> modeleBC
Call:
aov(formula = BC ~ activity + event, data = dfVariance)

Terms:
              activity              event      Residuals
Sum of Squares 6.026018e+09 1.011321e+09 9.922996e+13
Deg. of Freedom          9              9          86518

Residual standard error: 33866.33
Estimated effects may be unbalanced
68765 observations deleted due to missingness
> summary(modeleBC)
              Df      Sum Sq   Mean Sq F value Pr(>F)
activity       9 6.026e+09 6.696e+08   0.584  0.812
event          9 1.011e+09 1.124e+08   0.098  1.000
Residuals    86518 9.923e+13 1.147e+09
68765 observations deleted due to missingness
> |

```

Annexe 9 : Analyse de variance de polluant BC

```

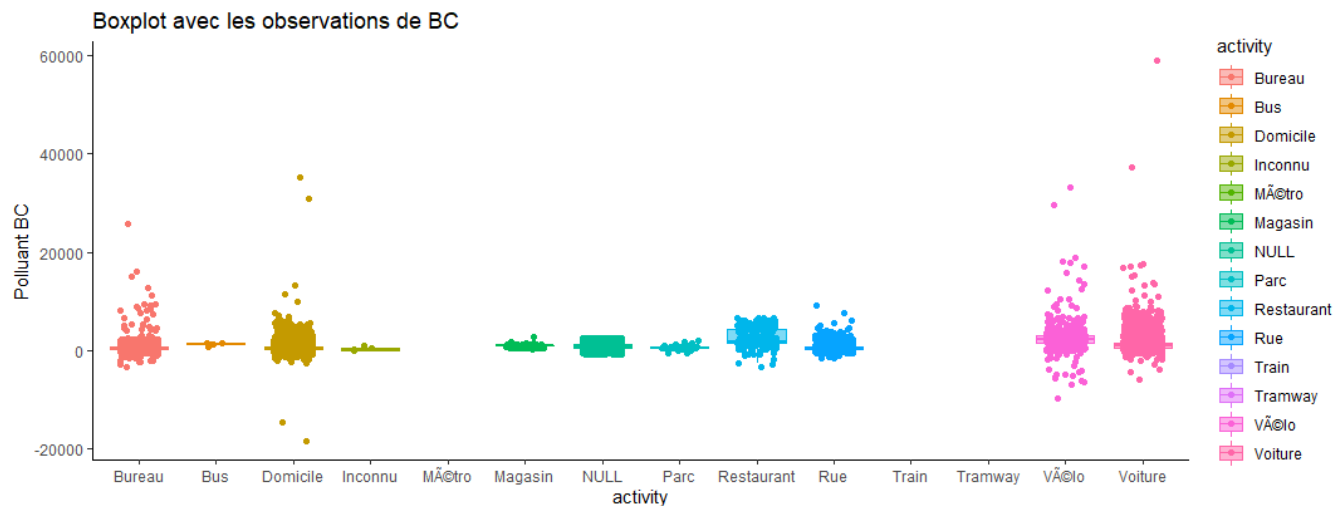
> modelePM1.0
Call:
aov(formula = PM1.0 ~ activity + event, data = dfVariance)

Terms:
              activity              event Residuals
Sum of Squares    751908    180387 11068981
Deg. of Freedom     12              9   155280

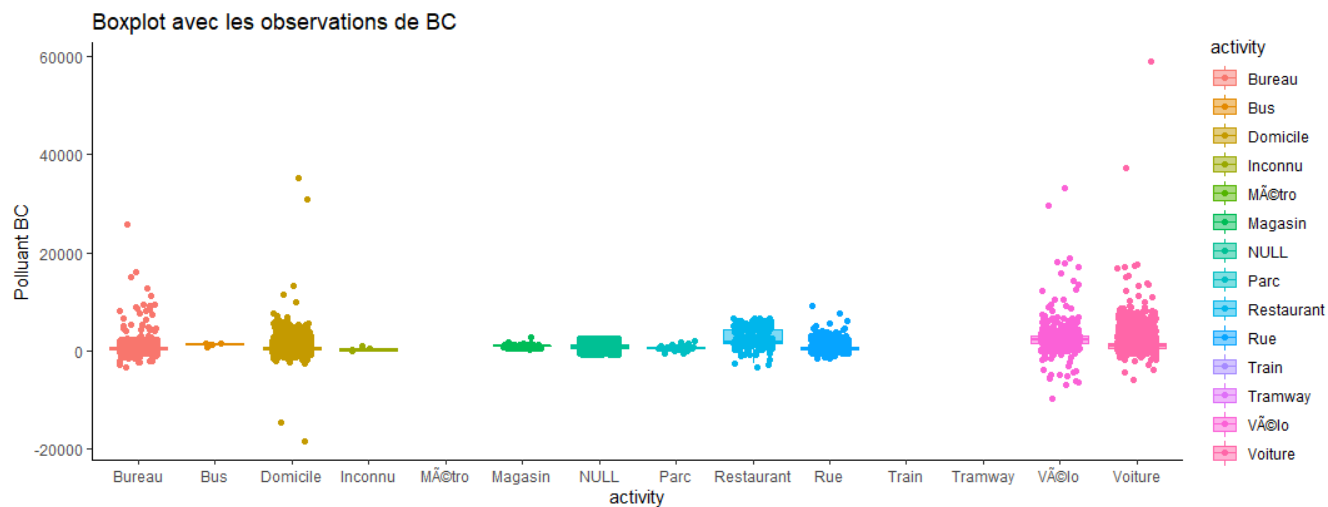
Residual standard error: 8.442986
Estimated effects may be unbalanced
> summary(modelePM1.0)
              Df      Sum Sq Mean Sq F value Pr(>F)
activity       12    751908   62659   879.0 <2e-16 ***
event          9    180387   20043   281.2 <2e-16 ***
Residuals    155280 11068981       71
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>

```

Annexe 10 : Analyse de variance de polluant PM1.0



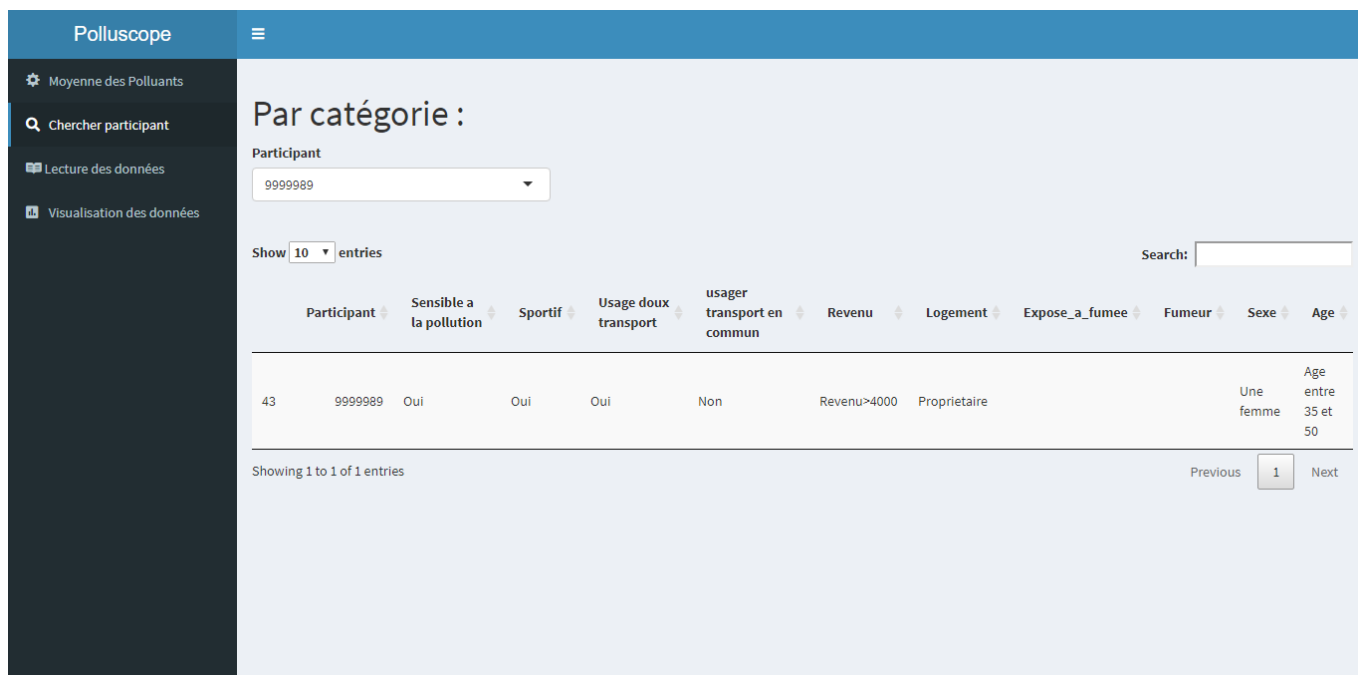
Annexe 11 : Les outliers de BC avant traitement



Annexe 12 : Les outliers de BC apr s traitement



Annexe 13 : L'interface pour visualiser les moyennes des polluants par catégorie



Annexe 14 : L'interface pour chercher les données d'un participant