Les aménagements cyclables et la sécurité

Yasmine Maricar – Emmnauel Bavoux – Thomas Nguyen – Guillaume Lebault
Datamining _ M2 IMSD 2016/2017

université PARIS-SACLAY

1. INTRODUCTION

Dans le cadre du projet « Paris, 2020, capitale du vélo », Paris se fixe l'objectif de devenir la capitale mondiale du vélo avec l'objectif d'atteindre 15% de déplacement effectués à vélo d'ici 2020. L'ambition est forte au regard de la culture cycliste présente dans les pays voisins, en Allemagne et au Pays-Bas. Amsterdam, Berlin, Munich et tant d'autres encore ont intégré les vélos au cœur de la ville en fournissant des espaces de circulation aménagé. Les parkings à vélos, en autre, où les vélos rangés se comptent par dizaines voire centaines donnent un aperçu de l'imprégnation du vélo dans les modes de déplacement de ces pays.

Le cycliste parisien occupe certes une place plus importante qu'auparavant. Et les parisiens sont de plus en plus en plus nombreux à utiliser le vélo comme un moyen de transport alternatif à la voiture ou aux transports en communs. La mise en place du Vélib' et la multiplication des aménagements cycliste ont permis cet essor. Ces dispositifs prennent la forme de piste cyclable, séparée ou non de la chaussée, de l'accès à des couloirs de bus, de voie à contresens dans certaines rues. Ces éléments n'ont pas pour seul rôle de faciliter la circulation, il permette également d'assurer la sécurité des vélos en séparant le flux de voitures de celui des vélos.

Alors que la Mairie de Paris prévoit d'agrandir le réseau cycliste parisien « de 700 km à 1 400 km », une revue des aménagements en place et de la sécurité routière des cyclistes paraît nécessaire. Cet audit doit permettre d'une part de décrire la situation actuelle et d'autre part de formuler des recommandations sur la mise en place de nouveaux aménagements. C'est dans ce contexte *hypothétique* que se place notre projet de Data Mining. L'équipe ayant travaillé sur ce projet est constitué de :

- Guillaume LEBAULT
- Thomas NGUYEN
- Yasmine MARICAR
- Emmanuel BAVOUX

2. DEMARCHE D'ANALYSE

Dans l'objectif d'évaluer les politiques d'aménagements cyclistes mis en place à Paris au cours des dernières années, une démarche d'analyse structurée a été établie. Elle doit permettre dans un premier temps de dresse une typologie des accidents impliquant les vélos. Deuxièmement, elle propose un inventaire des aménagements cyclistes existants. Enfin, elle propose une vision croisée afin d'évaluer la pertinence de chaque type d'aménagement en termes de sécurité.

D'une manière générale, l'utilisation des bases de données disponibles donne lieu à une vision par accident et à une vision par lieux. Ces deux visions sont explorées successivement.

Procédure d'analyse de données

Nous avons décidé de structure notre analyse de données afin de garantir la solidité de nos résultats. Les étapes suivantes sont faites :

- Nettoyage des données, recodage de variables et création d'une base de données à analyser.
- Réaliser des statistiques descriptives. Les valeurs manquantes et les biais potentiels de chaque variable sont étudiés à ce stage.
- Produire des analyser statistiques permettant de dresser une typologie des accidents.
- Modéliser les données afin de tenir compte des interactions multiples entre les variables.

Données utilisées

Trois bases données ont été mobilisées pour mener à bien cette analyse.

- La base nationale des accidents en 2015 : Elle liste l'ensemble des accidents ayant eu lieu au cours de l'année 2015, précise les caractéristiques de l'accidents, les usagers et les véhicules impliqués.
- La base de aménagements cyclistes en 2015 : Elle liste les aménagements cyclistes en place en 2015 par portion continue d'aménagement, ce qui implique l'existence possible de plusieurs aménagements au sein d'une même rue. Une typologie des ces aménagements est proposé au sein de la base de données et permet de la distinguer.
- La base nationale d'adresse ouverte (BANO) : Elle offre une liste non-exhaustive des adresses parisiennes ainsi que des coordonnées GPS qui s'y rattache.

La base nationale des accidents

Chaque correspond à l'enregistrement d'un accident auprès des services du Ministère de l'Intérieur. Ces évènements sont datés, notre unité d'observation est donc un évènement temporel et géographiquement identifié par une adresse. La source de ces enregistrements est le constat d'accident rempli pour chaque accident. Les données rentrées de manière manuelle sont donc susceptibles de comporter des erreurs et des approximations. La base de données mise en ligne sur *OpenDataGouv* est fournie telle quelle et sans retraitement préalable. *La* base de données contient quatre tables :

- Une table « caractéristiques » : les caractéristiques générales de l'accident sont identifiées par un numéro d'accident
- Une table « lieux » : le lieu de l'accident est identifié par un numéro d'accident
- Une table « véhicule » : chaque véhicule impliqué dans l'accident, identifié par un numéro d'accident et de véhicule
- Une table « usagers » : chaque usager impliqué dans l'accident, identifié par un numéro usager et un numéro véhicule

L'enjeu est donc d'extraire les informations nécessaires pour chaque accident des deux dernières tables afin d'obtenir une ligne par accident.

Aperçu de données

Nous disposons de plus de 700 observations pour les accidents de vélos incluant des cyclistes sur l'année 2015 à Paris sur un ensemble de 6153 accidents. La *table 1* présente un aperçu des données.

Table : Premières variables et premières observations

	Num_Acc	mois	jour	lum	agg	int	atm	col
49730	201500050509	4	5	1	2	1	1	3
49731	201500050510	4	9	1	2	5	1	3
49739	201500050518	4	25	1	2	1	1	3
49740	201500050519	4	25	1	2	2	2	3
49753	201500050532	5	19	1	2	2	1	7

Enjeux

Une analyse préliminaire des données permet de dégager les observations suivantes :

- La localisation des accidents est imprécise car le numéro d'adresse est parfois manquant. L'orthographe de l'adresse n'est également par toujours correcte. Des retraitements sont nécessaires.
- L'heure des accidents est approximative car des points de masse sont observées. Les heures pleines et les quarts d'heures sont plus fréquents. Dans le contexte d'un accident il apparaît donc que les personnes qui remplissent le constat
- Des données sont manquantes sur certaines variables qui caractérisent les accidents.
- Les conditions de circulation changent de jours en jours (ex. : présence de travaux).

La base nationale des accidents

Presque prête à l'emploi, l'enjeu est d'utiliser les données localisées pour agrémenter nos données sur les caractéristiques des accidents :

- Comment croiser cette base avec la base d'accidents de la route.
- Proposer des outils de visualisation des accidents.

La base nationale d'adresse ouverte (BANO)

La base nationale d'adresse consiste en une liste de rue avec les coordonées GPS correspondantes.

Méthodologie pour l'analyse de données qualitatives

Les bases disponibles sont essentiellement composées de données qualitatives, c'est-à-dire proposant des caractéristiques pour lequel un mesure tel qu'une longueur ou un ordre n'a pas de sens. Cependant la plupart de ces variables sont modales ; elles ne prennent qu'un nombre limité de valeurs. *De facto*, il est nécessaire d'adopter des outils statistiques adaptés. Par ailleurs, si les données qualitatives sont moins susceptibles aux approximations, elles sont plus sensibles à la présence de données manquantes.

Parmi les outils statistiques qui existent, le « V de Cramer » permet d'analyser le lien entre deux variables qualitatives modales. Concrètement, en observant les fréquences croisées il est possible de détecter des associations entre deux modalités. Elles peuvent par exemple se produire simultanément. Cette analyse de contingence synthétiser par le « V de Cramer » correspond à la corrélation des variables quantitatives. Il varie entre 0 et 1.

Valeur	Force du lien statistique
0	Absence de relation
Entre 0,05 et 0,10	Très faible
Entre 0,10 et 0,20	Faible
Entre 0,20 et 0,40	Modérée
Entre 0,40 et 0,80	Forte
Entre 0.80 et 1	Louche (Colinéarité)

3. ACCIDENTOLOGIE SUR PARIS

Une vision de l'ensemble des accidents est nécessaire afin de bien comprendre la nature d'un accident de la route à Paris. Une attention particulière sera portée à dresser les caractéristiques des accidents impliquant au moins un cycliste. La base « accidents » du Ministère de l'Intérieur ayant été fournie sans modifications préalables, des retraitements préliminaires sont nécessaires.

L'objectif est de constituer **une table unique d'accidentologie dont l'unité d'observation est l'accident.** Il est donc nécessaire de croiser les quatre tables disponibles au sein de la base « accidents ».

Retraitements préliminaires

Aucune modification majeure n'a été faite sur la structure des tables « caractéristiques » et « lieux » puisque la référence de l'accident peut être utilisée comme index principal.

En revanche, dans le cas des tables « véhicules » et « usagers », la référence de l'accident ne peut être utilisée en index principal puisqu'un accident peut concerner plusieurs véhicules et plusieurs usagés. Par conséquent, la référence d'un accident peut être identique d'une observation à une autre.

Recodage des variables

Principe

Les tables « véhicules » et « usagers » contiennent des variables propres à chaque véhicule et chaque usager présent lors d'un accident.

Dans le but de fusionner les observations ayant la même référence d'accident, ces variables, qui sont qualitatives, ont été recodées de façon à ce qu'elles soient quantitatives en appliquant la méthode suivante :

- Regroupement si nécessaire des modalités pour limiter leur nombre et ainsi le nombre de variable à créer
- Construction de nouvelles variables représentant la somme des modalités présentes dans un même accident.

Ce recodage et cette restructuration de la table nous permet alors de fusionner les observations ayant les mêmes références d'accident.

Table véhicules

Les variables « catv » (= catégorie de véhicule) et « manv » (= manœuvre précédent l'accident) sont propres à chaque véhicule et justifient qu'il y ait plusieurs observations pour une même référence d'accident.

Recodage de « catv » :

« catv » comprend une vingtaine de modalités qui sont regroupées en 5 modalités : véhicule ordinaire, véhicule lourd, deux roues, vélo et autres.

Pour chaque modalité les variables représentant le nombre de véhicules ordinaires, lourds concernés par l'accident, sont construites.

Recodage de « manv » :

Les variables représentant le nombre de chaque type de manœuvre précédant l'accident sont construites.

On peut alors fusionner les observations présentant une référence d'accident commune.

Table 2 : Table avec les variables de manœuvre et de catégorie de véhicule recodées

	Num_Acc	manAutre	manDepa	manTour	manDep	nbrAutres	nbrVelo	nbrVehiL	nbrVehiO	obsm
58654	201500058654	2	0	0	0	0	0	0	1	2
58653	201500058653	1	0	0	0	0	0	0	1	1
58652	201500058652	1	1	0	0	0	0	0	1	2

Table usagers

Les variables « grav » (= gravité de l'état de l'accidenté), « secu » (= équipement de sécurité) et « an_nais » (=date de naissance) sont propres à chaque usager et justifient que qu'il y ait plusieurs observations à une référence d'accident.

La méthode utilisée pour le recodage est similaire à celle utilisé pour la table « véhicules ».

Recodage de la variable « grav »

Les variables représentant le nombre de morts, hospitalisés, blessés légers et d'accidentés indemnes sont construites.

Recodage de la variable « sécu »

Seules les modalités correspondant à l'utilisation d'un casque ou d'un équipement réfléchissant sont comptés.

Les variables représentant le nombre d'usagers utilisant un casque et le nombre d'usagers utilisant un équipement réfléchissant dans un même accident ont été intégrées.

Recodage de la variable « an_nais »

L'année de naissance a été convertie en âge et a été ensuite regroupée par tranche : 0 – 15 ans, 15-25 ans, 25 – 60 ans et plus de 60 ans.

Les individus dans un même accident partageant la même tranche d'âge ont été ensuite sommés dans leur variable respective.

Table 3 : Table « usagers » après recodage des variables

	Num_Acc	a_60P	a_60	a_25	a_15	u_r	u_c	g_T	g_H	g_B	g_I
57673	201500058654	0	2	0	0	0	1	0	0	1	1
57672	201500058653	0	0	1	0	0	0	0	0	0	1

Construction de la table de travail

Le recodage des variables énoncées ci-dessus et la restructuration des tables « véhicules » et « usagers » ont permis d'obtenir deux tables à identifiant d'accident unique.

Puisque cet identifiant d'accident unique est partagé par les quatre tables, en concaténant les variables selon ce dernier, une table regroupant l'ensemble des variables fournies a pu être crée. Elle peut alors servir de base de travail pour les analyses.

Description des accidents

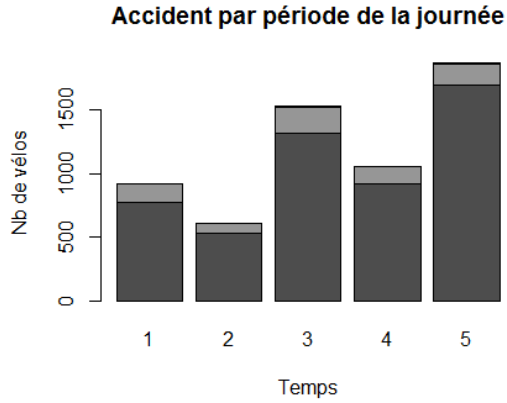
Nous allons étudier les distributions des modalités en fonction des véhicules dans la base de données finale recodée dans la partie précédente (« Base nationale des accidents »).

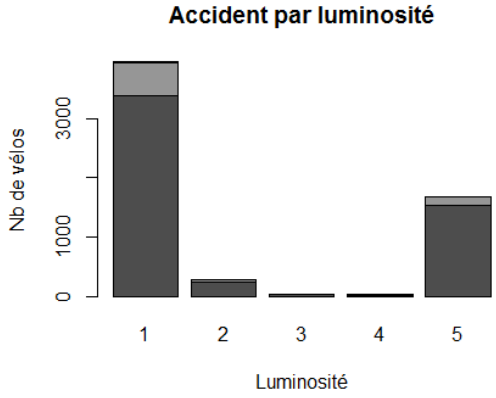
Notre table finale des observations d'accidents corporels liés à la circulation contient les variables :

"Num_Acc", "mois", "jour", "lum", "agg", "int", "atm", "col", "adr", "circ", "nbv", "prof", "plan", "larrout", "surf","infra", "manAutre", "manDepa", "manTour", "manDep","nbrAutres", "nbrVelo","nbrVehiL", "nbrVehiO", "obsm", "a_60P", "a_60","a_25","a_15", "u_r", "u_c", "g_T", "g_H", "g_B", "g_I", "temps"

Fréquences croisées

Caractéristiques diverses

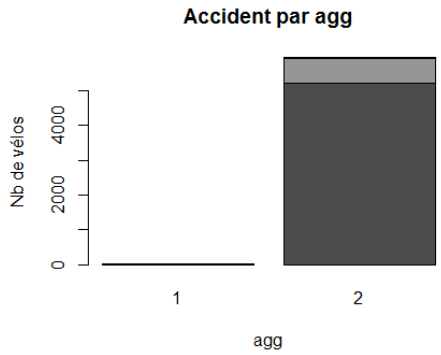




	1	2	3	4	5
0	774	534	1319	920	1693
1	143	77	203	133	171
2	4	1	2	5	3
3	0	0	1	0	0

	1	2	3	4	5
0	3398	246	36	22	1538
1	548	38	2	5	134
2	14	0	0	0	1
3	1	0	0	0	0

Les accidents se produisent plus souvent en plein jour (lum=1) et dans une moindre mesure, la nuit (lum=5).

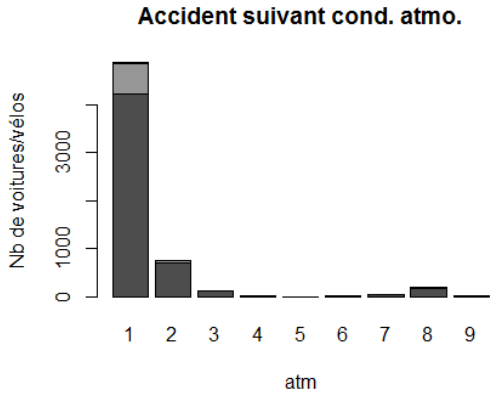


	1	2
0	30	5210
1	1	726
2	0	15
3	0	1

Les accidents se produisent en majorité en agglomération.

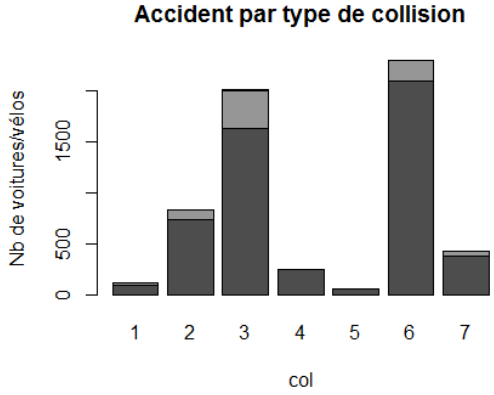
	1	2	3	4	5	6	7
0	2093	1229	688	326	425	8	471
1	250	199	103	22	65	1	87
2	6	2	6	0	1	0	0
3	1	0	0	0	0	0	0

Les accidents se produisent le plus hors-intersection, intersection en X et intersection en T.



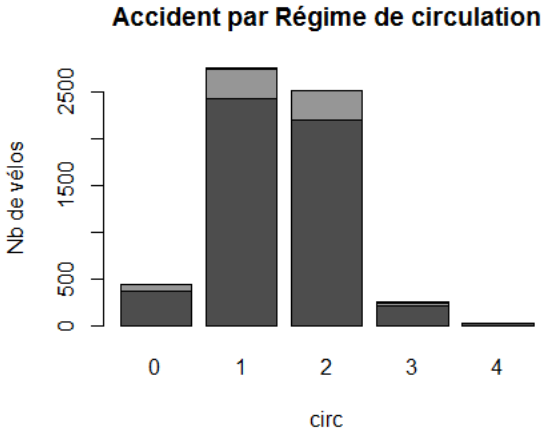
	1	2	3	4	5	6	7	8	9
0	4217	688	116	7	1	4	32	162	13
1	638	61	2	0	0	0	6	20	0
2	15	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0

Les accidents surviennent le plus souvent en condition atmosphérique « Normale » puis « Pluie légère », « Temps couvert » et « Pluie forte ».



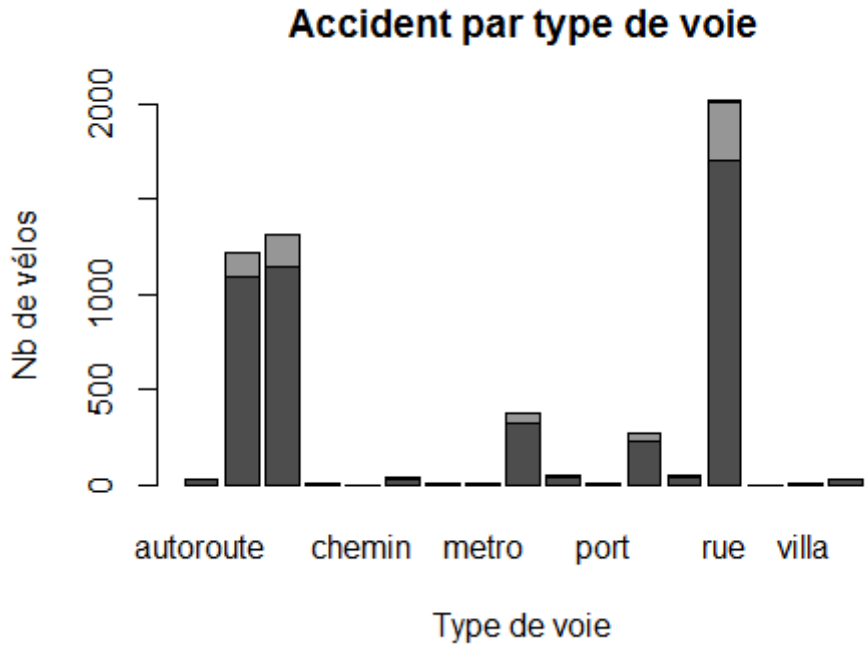
	1	2	3	4	5	6	7
0	90	733	1638	245	51	2106	377
1	20	96	367	5	3	192	44
2	2	2	8	0	0	3	0
3	0	0	0	1	0	0	0

Nous sommes souvent confrontés à des cas de collisions de deux véhicules (par le côté (modalité 3, plus importante concernant les vélos) ou par l'arrière). La modalité 6 regroupe les collisions regroupées sous « autre collision » et concentre la majorité de nos observations.



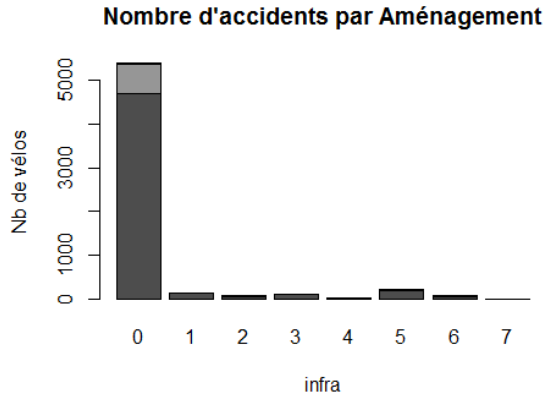
	0	1	2	3	4
0	368	2436	2201	217	18
1	70	314	309	28	6
2	1	6	7	1	0
3	0	0	1	0	0

Les accidents se produisent le plus souvent sur les voies à sens unique ou bi-directionnelles.



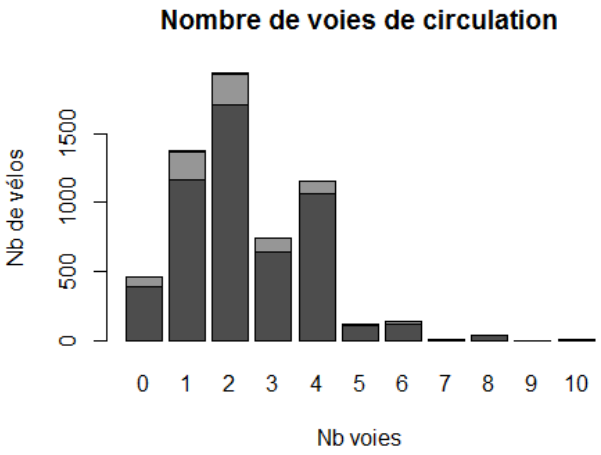
Le type de voie qui se démarque concernant les accidents corporels de la circulation sont les rues, les boulevards et les avenues.

	autoroute	avenue	boulevard	cours	place	pont	quai	route	rue	voie
0	28	1094	1140	32	327	39	224	34	1701	24
1	0	123	171	3	53	8	41	8	308	1
2	0	1	3	0	0	1	2	2	6	0
3	0	0	0	0	0	0	0	1	0	0



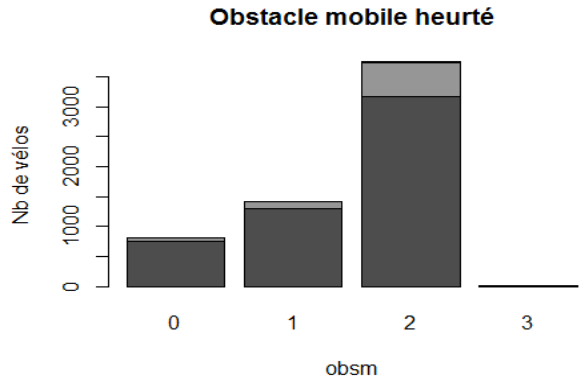
	0	1	2	3	4	5	6
0	4707	126	60	97	5	185	59
1	676	3	12	1	0	25	9
2	13	0	0	0	0	0	0
3	1	0	0	0	0	0	0

Le plus souvent, nous ne savons pas de quel type d'aménagement il s'agit. La modalité la plus Importante ensuite sont les carrefours aménagés.



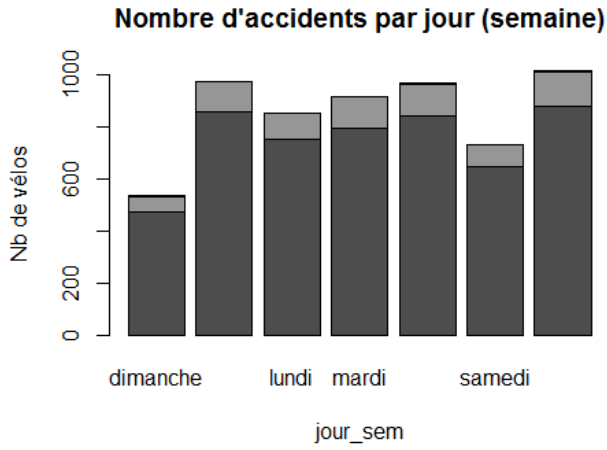
Le nombre de voies ne semble pas avoir une relation linéaire avec les accidents.

Les accidents surviennent le plus souvent sur des lieux avec 2 voies/1 voie de circulations.

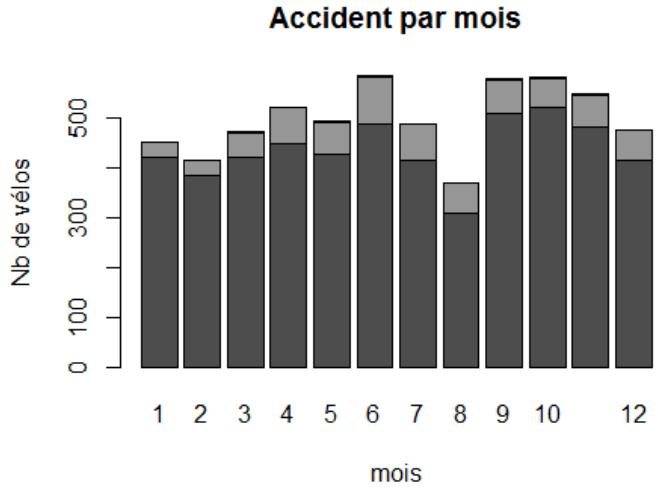


	0	1	2	3
0	761	1302	3166	11
1	55	108	563	1
2	0	0	15	0
3	0	0	1	0

Le plus souvent, c'est un autre véhicule qui est heurté (modalité 2) ou bien un piéton (modalité 1).



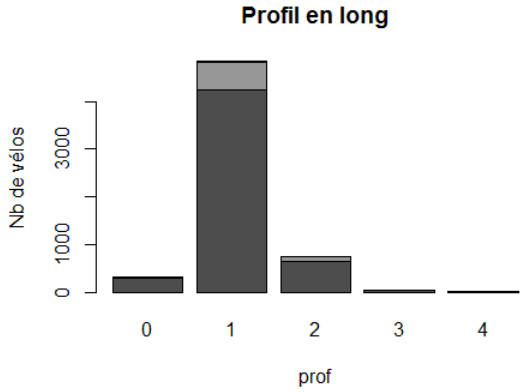
	dimanche	jeudi	lundi	mardi	mercredi	samedi	vendredi
0	474	857	749	793	841	648	878
1	58	113	103	121	119	81	132
2	1	2	0	1	6	2	3
3	1	0	0	0	0	0	0

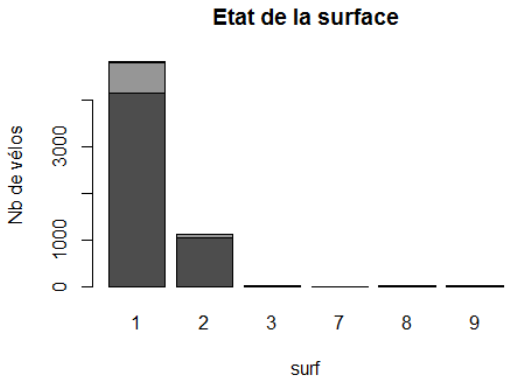


On remarque que le mois de Juin concentre le plus d'accidents et qu'il y en a le moins en Août concernant les véhicules hors vélo. Concernant les jours de l'année, les accidents semblent uniformément répartis dans le mois.

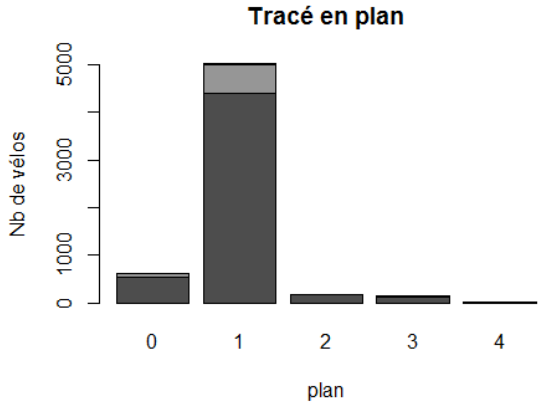
	1	2	3	4	5	6	7	8	9	10	11	12
0	422	385	421	447	426	489	415	309	510	522	480	414
1	30	31	49	75	66	93	73	59	66	57	66	62
2	0	0	2	0	2	2	1	0	2	4	2	0
3	0	0	0	0	0	0	0	1	0	0	0	0

Caractéristiques de la route (prof, plan et surf)





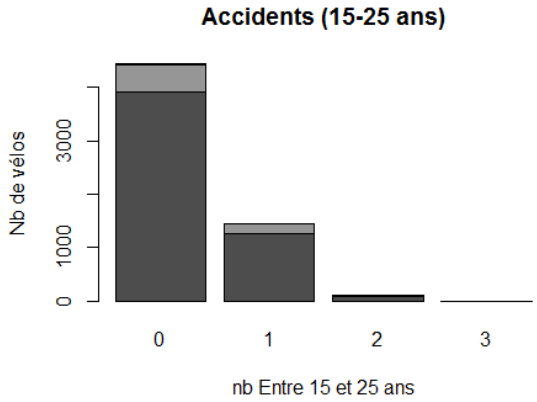
Le profil en long est en grande majoritairement plat. L'état de la surface est normal ou bien mouillé pour la grande majorité des accidents survenus.



Le tracé en plan est généralement rectiligne.

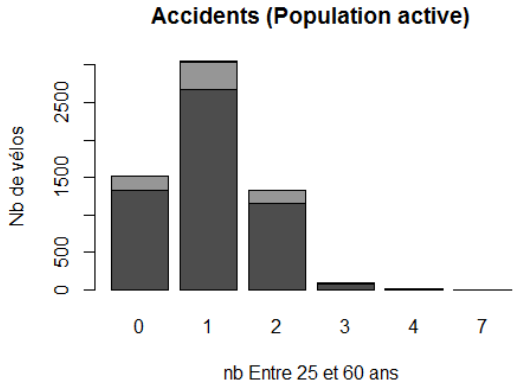
Caractéristiques des usagers concernés par accident (grav, age, sécu)

Les personnes de – de 15 ans ne sont impliqués que dans 27 accidents (1 personne concernée) parmi notre base de travail de 5983 observations d'accidents.



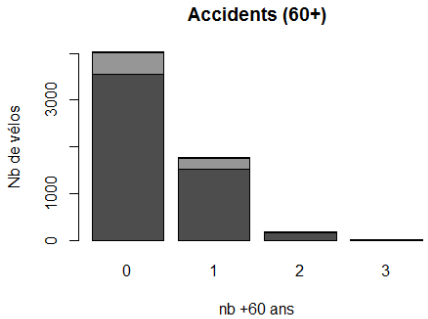
	0	1	2	3
0	3910	1255	74	1
1	519	186	22	0
2	11	2	2	0
3	0	1	0	0

Les 15-25 ans sont concernés dans 25% des cas d'accidents et concentrent près de 30% des cas d'accidents impliquant au moins un vélo.



	0	1	2	3	4	7
0	1327	2673	1149	81	9	1
1	183	372	171	1	0	0
2	2	6	7	0	0	0
3	0	1	0	0	0	0

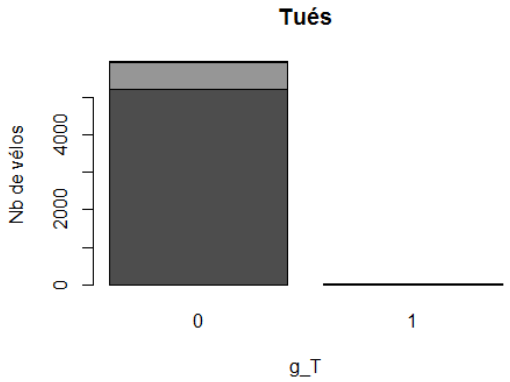
74% des accidents touche la population active, de plus, on retrouve au moins 2+ personnes de cette tranche d'âge impliquée dans 24% des cas. Enfin, ils représentent les trois quarts des accidents impliquant des vélos.

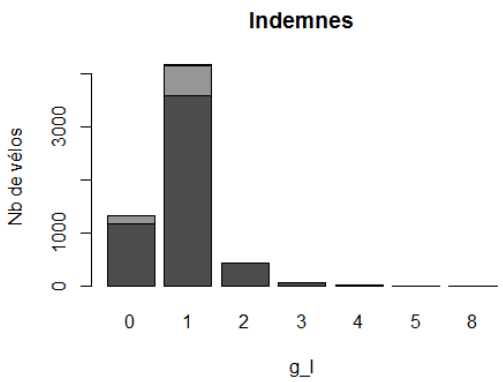


Les 60+ sont présents dans 30% des accidents et 30% des accidents impliquant un vélo ou plus.

	0	1	2	3
0	3547	1525	158	10
1	469	233	25	0
2	11	4	0	0
3	0	1	0	0

Gravité des accidents





Dans notre échantillon d'étude, l'année 2015 a compté 25 tués par accident. Au moins une personne s'en sort indemne dans 78% des cas mais au moins 1 personne est blessée légèrement dans 65% des cas. Quant aux blessés hospitalisés, ils représentent 7% des effectifs. Les cyclistes semblent principalement s'en sortir indemnes ou blessés légers.

	0	1
0	5221	19
1	722	5
2	15	0
3	1	0

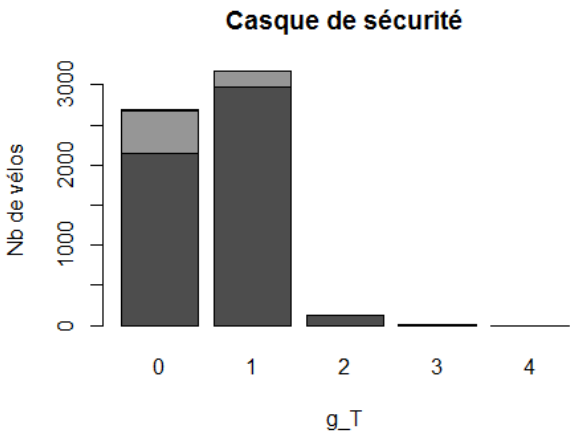
	0	1	2	3	4	5	8
0	1156	3587	419	66	10	1	1
1	152	565	10	0	0	0	0
2	5	9	1	0	0	0	0
3	1	0	0	0	0	0	0

	0	1	2	3
0	1908	3204	121	7
1	145	560	22	0
2	2	10	3	0
3	0	0	0	1

	0	1	2
0	4862	373	5
1	675	52	0
2	12	3	0
3	1	0	0

Utilisation d'un casque ou d'un équipement réfléchissant

Aucune personne dans la base extraite que l'on étudie n'utilisait un équipement réfléchissant.



On remarque que la majorité des cyclistes ne portait pas de casques.

	0	1	2	3	4
0	2139	2979	119	2	1
1	531	186	10	0	0
2	13	2	0	0	0
3	0	0	0	1	0

ACM sur la base « table_finale_totale.csv »

L'analyse des correspondances multiples (ACM) permet d'analyser un nombre important de variables dites « qualitatives » ce qui est le cas de notre jeu de données initial qui en est presque exclusivement pourvu.

Dans notre cas, nous utilisons la fonction **MCA** du package **FactoMineR**.

L'idée de cette analyse est de réduire les dimensions en essayant d'identifier les modalités actives les plus significatives caractérisant le mieux un accident au sens large du terme.

Le jour de la semaine a-t-il de l'importance ? Ou est-ce-plutôt le mois ou encore les conditions de la chaussée ?

Notre démarche commence par choisir les variables que nous considérons utiles dans l'analyse en créant un jeu de données ne conservant que celles-ci :

mois	infra	a_60P
lum	manAutre	a_60
int	manDepa	a_25
atm	manTour	a_15
col	manDep	g_T
circ	nbrAutres	g_H
nbv	nbrVelo	g_B
prof	nbrVehiL	g_I
plan	nbrVehiO	temps
surf	obsm	jourSem

Notre ACM a été configurée pour ne conserver l'information que pour 5 axes avec l'argument **ncp=5**.

Lors de nos différentes manipulations, nous avons noté l'influence délétère des variables non renseignées sur le modèle ACM ; ainsi, nous avons supprimé les accidents pour lesquels nous avons des valeurs nulles ou pas de valeurs pour : prof, surf, plan, obsm, infra, nbv.

De même, grâce à un test V de Cramer, nous notons une forte corrélation entre les variables suivantes :

- **temp/lum => correlation Cramer 0.4 : nous gardons temp**
- **surf/atm => correlation Cramer 0.36 : nous gardons atm**

Nous passons de 5983 accidents observés à 4590.

Parmi nos variables sélectionnées, certaines sont quantitatives suite au recodage précédent de la base initiale :

man*, nbr*, a_*, g_*.

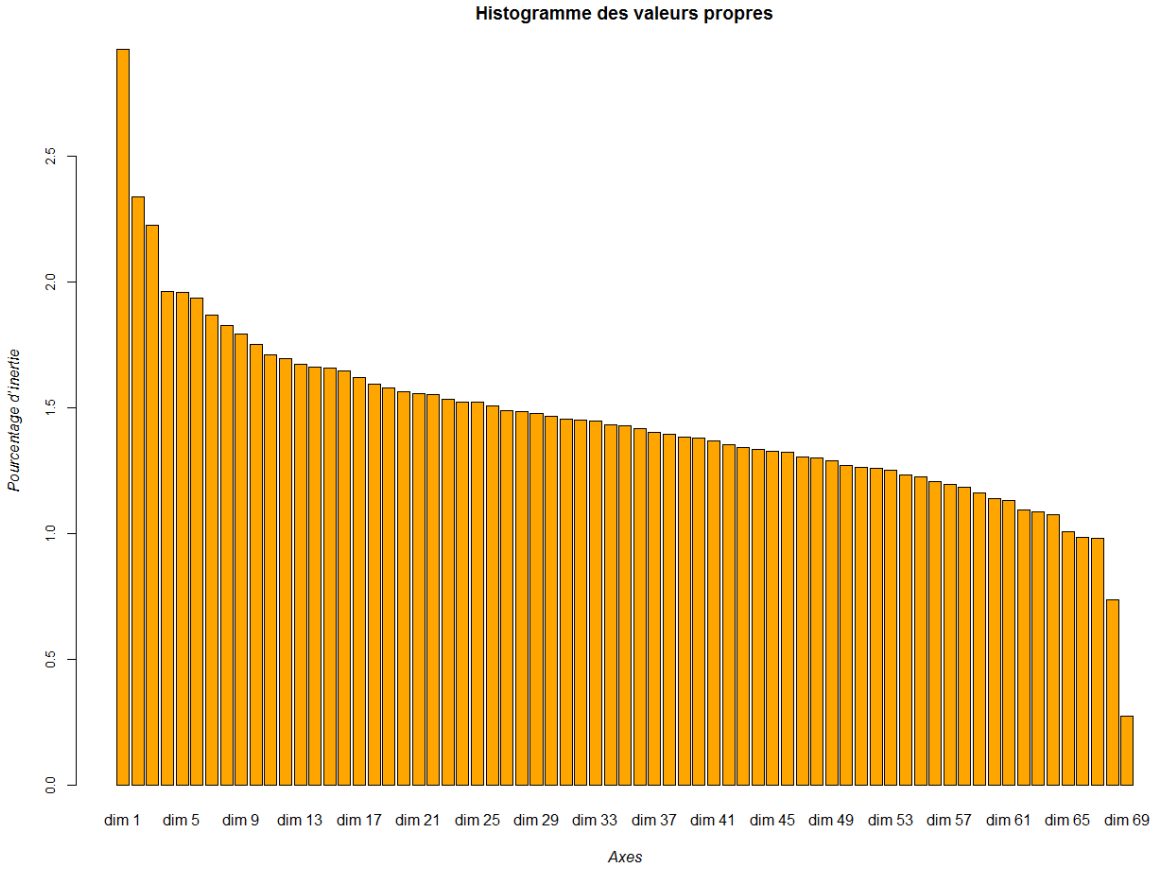
Elles sont exclus de l'ACM avec l'expression **quanti.sup=c(10:17,19:26)** dans notre code.

Une fois l'ACM créée en sortie dans **resultatmca**, 15 objets sont stockés ; nous allons nous intéresser aux objets 1 (valeurs propres : **$eig**), 3 (coordonnées des modalités actives : **varcoord**), 5 (contributions des modalités actives : **varcontrib**) et 8 (coordonnées des individus : **indcoord**).

Valeurs propres (eigen values)

Au regard du nombre de variables et du nombre de modalités, nous obtenons 83 facteurs, classés du plus important au moins important (ou du plus explicatif au moins explicatif). A chacun de ces 7 facteurs est associée une « valeur propre » qui traduit la quantité d'information expliquée par le facteur.

Le but de la démarche est de rendre compte de la plus grande partie possible de l'information originale avec le nombre le plus petit possible de « variables », l'examen de l'histogramme des valeurs propres va permettre de nous aiguiller quant au choix des facteurs à conserver.



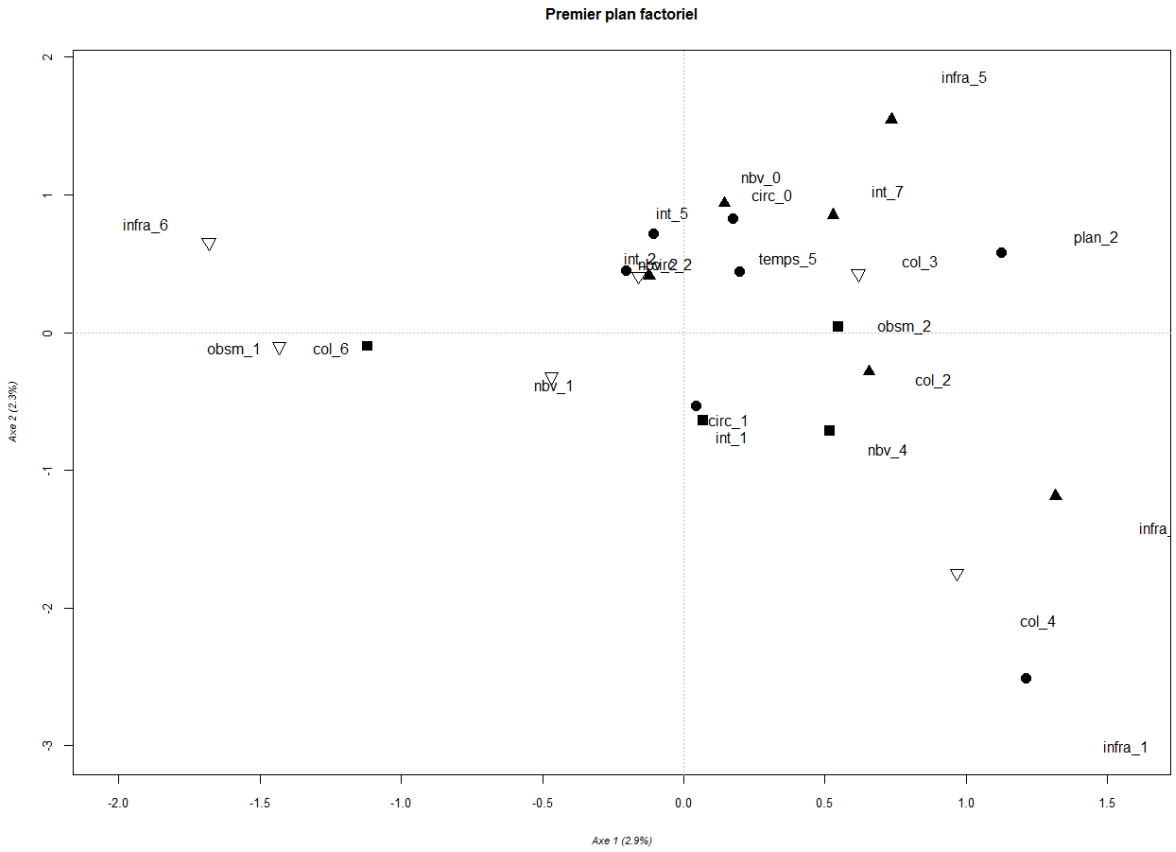
Visuellement, on se rend compte qu'il y a un facteur plus fort, les 15/20 premiers facteurs expliquant visuellement une partie de l'inertie.

Empiriquement, les facteurs dont la valeur propre est supérieure à la valeur propre moyenne (qui vaut par définition 1 divisé par le nombre de facteurs) sont conservés (critère de Kaiser). Ici, le critère de Kaiser inciterait à retenir les facteurs dont le pourcentage d'inertie est supérieur à (100/69=) 1,45%.

Les modalités actives à représenter

L'idée est de déterminer quelles sont les modalités actives que l'on souhaite représenter sur le graphique. Tout comme le critère de Kaiser pour les valeurs propres, retenir uniquement les modalités dont la contribution est supérieure à deux fois la contribution moyenne est une manière de faire (Cibois, 1986, 1997).

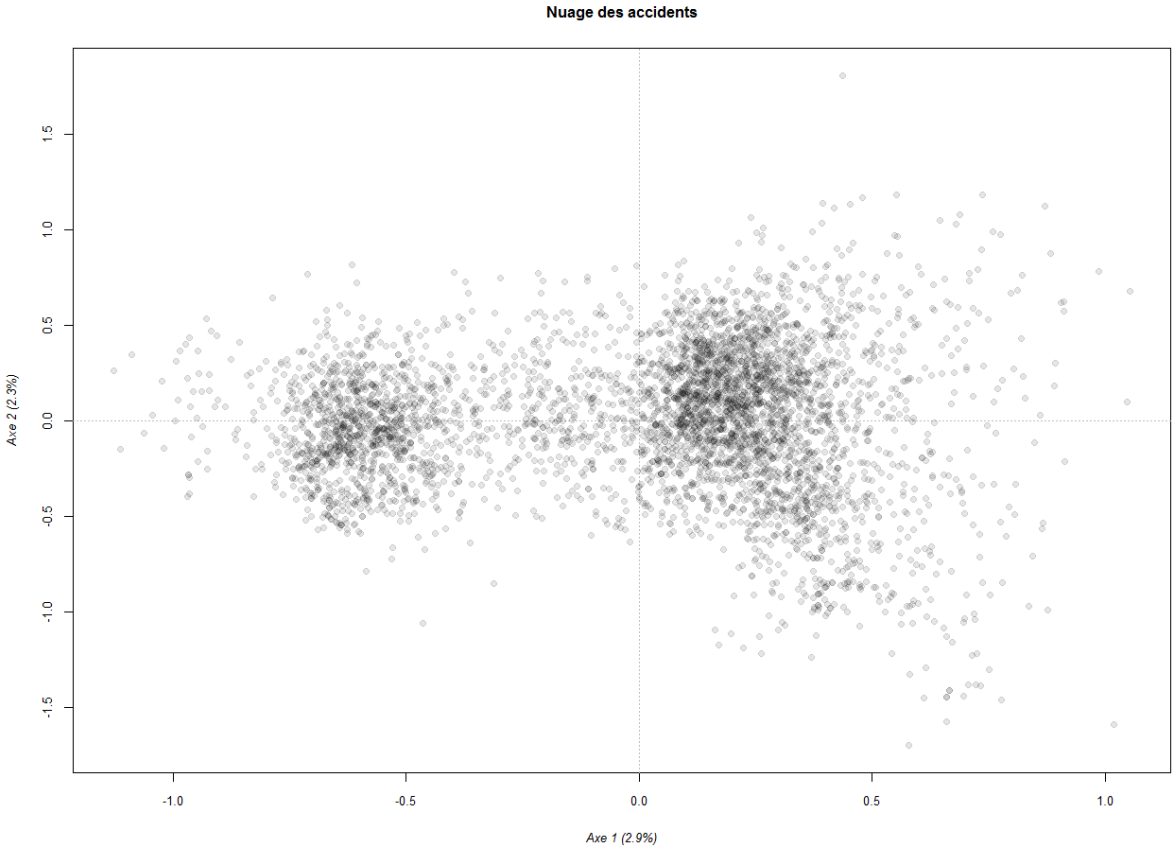
Notre code sélectionne donc les numéros de modalités remplissant cette condition afin de lancer une représentation graphique de celles-ci. Ces modalités sont retenues dans la table **moda** afin de les visualiser dans l'espace sur les 2 premiers axes.



Pour vérifier la pertinence de la répartition de ces modalités, nous pouvons comparer avec la répartition de nos observations, les accidents, dans l'espace.

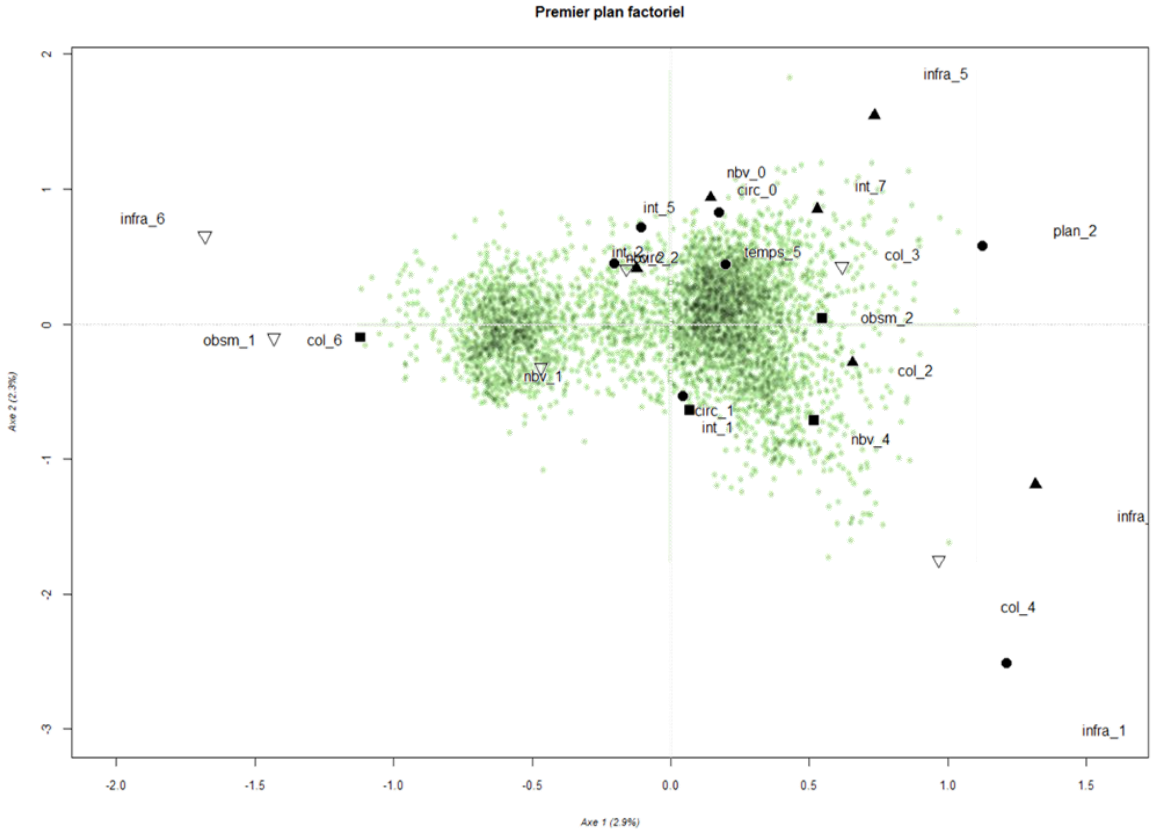
Le nuage des accidents

Le nuage des accidents va projeter dans l'espace toutes les observations, quelque soit le type de véhicule impliqué.



On distingue visuellement 2 groupes d'observations.

En superposant avec les modalités actives signifiantes :



Typologie des accidents de vélo

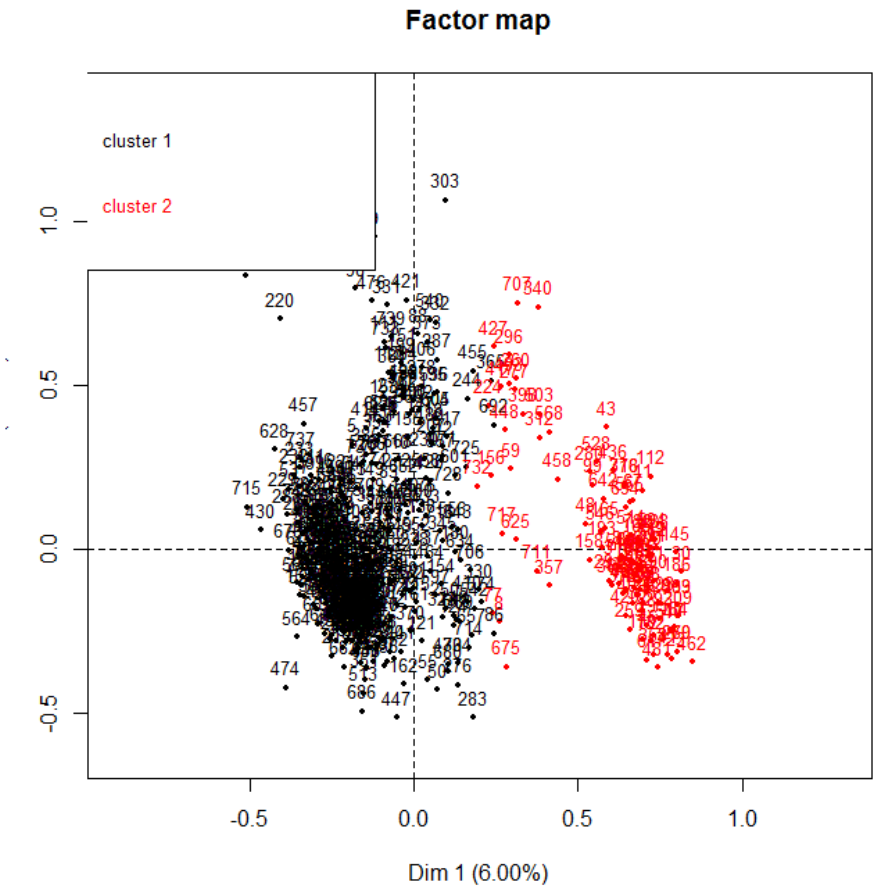
Une ACM suivant la même procédure que ci-dessus est appliquée aux seules observations de la table concernant au moins un vélo.

Les variables crées dans le traitement préliminaire ont été recodées en variables binaires. Elles représentent la présence d'une ou plusieurs éléments des variables dans l'accident.

Ceci permet inclure ces informations dans notre ACM en les recodant en variables qualitatives à 2 modalités : 0 et 1

Par exemple, si dans un même accident le nombre de véhicule lourd est égal à 2 alors la variable sera recodée en 1.

Afin de pouvoir distinguer différentes typologies d'accidents, nous analysons nos données par une Classification Hiérarchique sur Composantes Principales. L'analyse est faite grâce à la fonction HCPC du package FactoMineR.



Nous décidons de couper nos individus au plus haut possible pour avoir deux groupes les plus distincts possibles.

On voit alors que deux clusters sont mis en valeurs et les informations fournies par la fonction nous éclairent sur les compositions de ces derniers

```
$`1`
                      Cla/Mod    Mod/Cla    Global      p.value     v.test
obsm=obsm_2          94.89362 100.000000 84.990958 2.924857e-77  18.605038
nbrVehiO=nbrVehiO_1  98.91008  81.390135 66.365280 3.627371e-54  15.497130
g_B=g_B_1            93.95973  94.170404 80.831826 1.506606e-49  14.798108
col=col_3            99.34641  68.161435 55.334539 8.333634e-41  13.376166
a_60=a_60_1          86.01896  81.390135 76.311031 5.874352e-08   5.422584
manTour=manTour_1    94.73684  28.251121 24.050633 2.554497e-07   5.153661
a_60P=a_60P_1        91.13300  41.479821 36.708861 7.891055e-07   4.938043
manDep=manDep_1      98.59155  15.695067 12.839060 1.575978e-06   4.801352
col=col_2            97.46835  17.264574 14.285714 2.860163e-06   4.680615
.
.
.
manDep=manDep_0      78.00830  84.304933 87.160940 1.575978e-06  -4.801352
a_60P=a_60P_0        74.57143  58.520179 63.291139 7.891055e-07  -4.938043
manTour=manTour_0    76.19048  71.748879 75.949367 2.554497e-07  -5.153661
a_60=a_60_0          63.35878  18.609865 23.688969 5.874352e-08  -5.422584
g_B=g_B_0            24.52830   5.829596 19.168174 1.506606e-49 -14.798108
nbrVehiO=nbrVehiO_0  44.62366  18.609865 33.634720 3.627371e-54 -15.497130
col=col_6            28.57143   8.968610 25.316456 1.691396e-67 -17.358828
obsm=obsm_1           0.00000   0.000000 15.009042 2.924857e-77 -18.605038

$`2`
                        Cla/Mod    Mod/Cla    Global      p.value     v.test
obsm=obsm_1         100.0000000 77.5700935 15.009042 2.924857e-77  18.605038
col=col_6            71.4285714 93.4579439 25.316456 1.691396e-67  17.358828
nbrVehiO=nbrVehiO_0  55.3763441 96.2616822 33.634720 3.627371e-54  15.497130
g_B=g_B_0            75.4716981 74.7663551 19.168174 1.506606e-49  14.798108
a_60=a_60_0          36.6412214 44.8598131 23.688969 5.874352e-08   5.422584
manTour=manTour_0    23.8095238 93.4579439 75.949367 2.554497e-07   5.153661
a_60P=a_60P_0        25.4285714 83.1775701 63.291139 7.891055e-07   4.938043
manDep=manDep_0      21.9917012 99.0654206 87.160940 1.575978e-06   4.801352
g_I=g_I_0            36.7088608 27.1028037 14.285714 8.663935e-05   3.925251
plan=plan_1          20.6225681 99.0654206 92.947559 1.985216e-03   3.092435
.
.
.
g_I=g_I_1            16.4556962 72.8971963 85.714286 8.663935e-05  -3.925251
col=col_2             2.5316456  1.8691589 14.285714 2.860163e-06  -4.680615
manDep=manDep_1       1.4084507  0.9345794 12.839060 1.575978e-06  -4.801352
a_60P=a_60P_1         8.8669951 16.8224299 36.708861 7.891055e-07  -4.938043
manTour=manTour_1     5.2631579  6.5420561 24.050633 2.554497e-07  -5.153661
a_60=a_60_1          13.9810427 55.1401869 76.311031 5.874352e-08  -5.422584
col=col_3             0.6535948  1.8691589 55.334539 8.333634e-41 -13.376166
g_B=g_B_1             6.0402685 25.2336449 80.831826 1.506606e-49 -14.798108
nbrVehiO=nbrVehiO_1   1.0899183  3.7383178 66.365280 3.627371e-54 -15.497130
obsm=obsm_2           5.1063830 22.4299065 84.990958 2.924857e-77 -18.605038
```

Le cluster 1 est caractérisé par la collision avec un obstacle mobile, dans la majorité des cas une voiture. La blessure légère avec une collision sur le côté est aussi très caractéristique de ce groupe. Il concerne le plus souvent des individus âgés entre 25 et 60 ans et des manœuvres de tournant ou dépassement sont le plus souvent la cause de l'accident.

Le cluster 2 est signé par la collision avec un piéton, catégorisé ici comme « Autre collision », ces accidents n'impliquent donc pas de véhicules ordinaire et il n'y généralement pas de blessé légers.

4. AMENAGEMENTS CYCLISTES ET ACCIDENTS

Les aménagements cyclistes présents dans Paris en 2015 représentent 9 223 portions d'aménagements, ce qui représente 736 374 mètres d'aménagements. Parmi ces aménagements il faut distinguer cinq grandes catégories d'aménagements

- Des pistes cyclables et des bandes cyclables qui bénéfice d'une ligne continue ou d'une séparation nette de la zone de circulation des voitures. Elles peuvent être sur la chaussée ou bien placée sur le trottoir. Les pistes cyclables et les bandes cyclables sont dans leur définition similaire.
- Des couloirs de bus ouverts aux vélos permettent aux vélos de rouler sur une voie où seuls les bus et les vélos sont présents.
- Des continuités cyclables qui prennent la forme d'une ligne pointillée identifiant une zone pour les cyclistes mais sans séparation nette avec l'ensemble des autres véhicules.
- Des zones sans marquage et se composant de zones identifiées par un pictogramme cycliste. Elles sont soit des zones de transitions, intermédiaires entre deux aménagements cyclistes à une intersection, soit des autorisations de circulation cycliste à contresens. Ce dernier type constituant l'essentiel des aménagements cyclistes.

Graphique : répartition des aménagements cycliste par type en nombre de portions aménagées

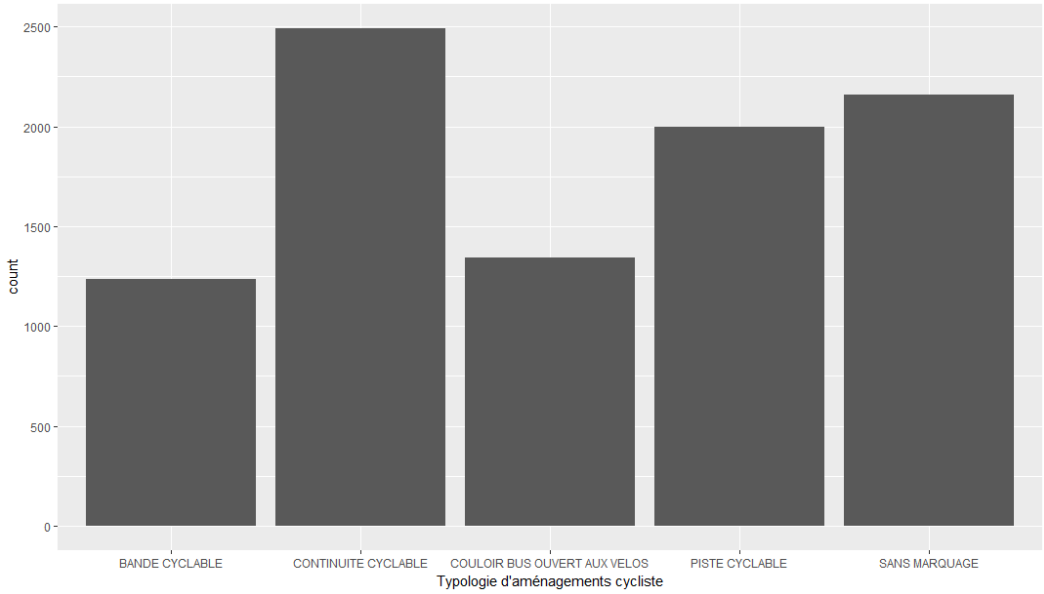


Tableau : aménagements cyclistes en fonction du sens de circulation en nombre de portions aménagées

TYPE DE VOIE / SENS DE CIRCULATION	N/A	CIRCULATION GENERALE INTERDITE	CONTRES ENS	SENS CIRCULATION GENERALE
BANDE CYCLABLE	4	1	419	812
CONTINUITE CYCLABLE	446	1	579	1464
COULOIR BUS OUVERT AUX VELOS	0	0	96	1245
PISTE CYCLABLE	68	5	528	1398
SANS MARQUAGE	110	176	1778	93

Base d'analyse et principe de modélisation

La présence d'information par portion de rue pour les aménagements cyclistes rend la lecture de ces aménagements en termes d'accidentologie difficile. En effet, les accidents sont identifiés par une adresse ou un nom de voie. Les aménagements à l'inverse sont présents pour un sens de circulation spécifique, dans une direction spécifique. Plusieurs aménagements peuvent également concurremment être présents en même temps au même endroit.

Afin de quantifier à la fois les accidents survenant dans une rue et les aménagements présents, une vision synthétique des aménagements a été extraite. En considérant l'accident comme un évènement ayant une probabilité faible de se produit à chaque instant *t*, on peut considérer que la taille de la rue importe : « plus l'on roule dans une rue, plus la probabilité d'avoir un accident est grande ». Cette simple mécanique de l'usage découle du fait qu'un cycliste à plus de chance d'avoir un accident à vélo plus son trajet est loin car les « opportunités d'accidents » sont plus grandes, *toutes choses égales par ailleurs*.

En suivant cette logique, l'information par portion d'aménagement cyclable peut être ramenée en calculant par rue la distance cumulée pour chaque type d'aménagement. Par ailleurs cette vision permet d'obtenir des données quantitatives. La distance est calculée à partir des tracés GPS qui sont fournies dans le champ « geo_shape » de la base d'aménagements cyclistes :

1. Conversion du champ geo_shape en un format JSON
2. Lecture du JSON et calcul pour chaque segment de la distance
3. Calcul de la distance totale.

L'ensemble des voies de Paris est ensuite constitué à l'aide de la base nationales d'adresses. Il est nécessaire de retraiter les adresses de chacune des bases « BANO », « accidents » et « aménagements » afin de séparer le numéro, le type de voie (« boulevard », « rue », etc.), le nom de la voie et l'indice de répétition (« bis », « ter » etc.) :

- Une comparaison de troncature est faite afin de recouvrer le nom de voie pour les accidents car beaucoup d'adresse apparaissent avoir été tronquées.
- Afin d'assurer une meilleure correspondance, un matching utilisant la distance de Levenshtein est réalisé.

Le nombre total d'accidents par rue est ensuite compté.

Statistiques descriptives

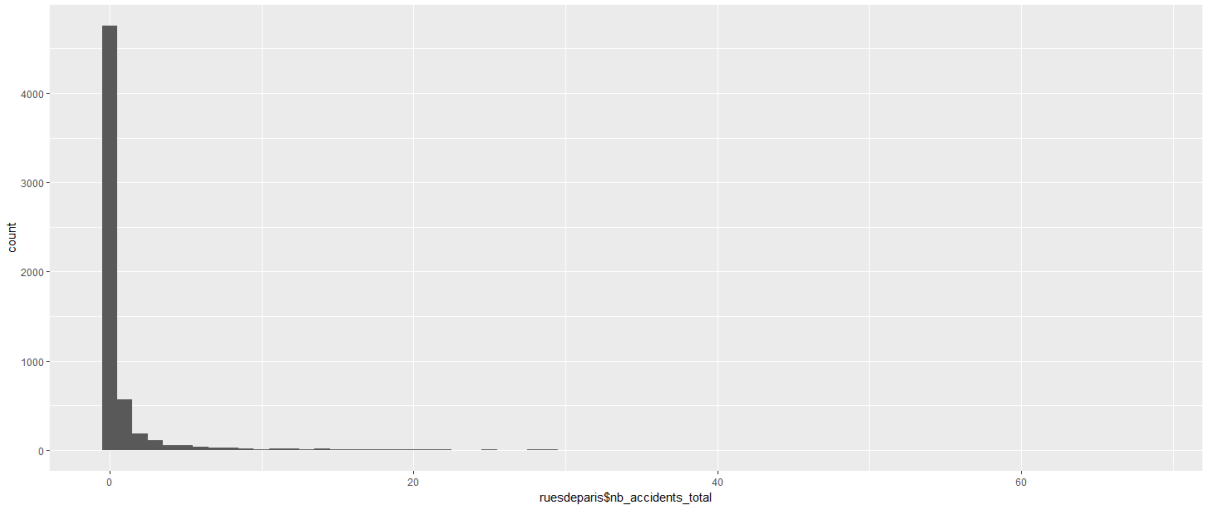
La base d'analyse se constitue des variables quantitatives suivantes :

Statistic	N	Mean	St. Dev.	Min	Max
longueur	6,005	199.407	304.313	0.000	4,221.278
l_couloirbus	6,005	20.515	134.792	0.000	3,511.153
l_sansmarquage	6,005	44.891	130.673	0.000	3,335.129
l_bandecyclable	6,005	13.589	92.710	0.000	3,190.903
l_pistecyclable	6,005	34.416	270.741	0.000	9,069.606
l_continuitecyclable	6,005	9.216	57.485	0.000	2,055.780
l_sens_general	6,005	59.024	310.139	0.000	7,466.382
l_sens_contresens	6,005	55.429	160.099	0.000	4,672.596
l_sens_na	6,005	4.870	62.110	0.000	3,382.410
l_sens_circInterdite	6,005	3.304	35.175	0.000	983.606
nb_accidents_total	6,005	0.996	3.911	0	68
nb_accidents_velo	6,005	0.124	0.655	0	17
nb_voies	6,005	1.415	0.748	0	6
nb_accidents_horsvelos	6,005	0.872	3.455	0	63

Les variables « l_ » sont des variables de longueur par rapport à une typologie d'aménagement ou un sens de circulation pour les variables « l_sens_ ». Les variables « nb_ » compte le nombre d'accidents totaux, nombre d'accidents à vélo, le nombre de voies, le nombre d'accidents hors vélos. La variable « longueur » corresponds à la longueur totale de la rue calculée à partir de la première et de la dernière adresse de la rue présente dans le BANO

De nombreuses rues ne sont le théâtre d'aucun ou de peu d'accidents. En conséquence, on observe des points de masses sur les faibles nombre d'accidents à vélo, à la fois pour les accidents à vélo et pour les accidents totaux. Un accident est en effet un évènement rare. Une observation graphique permet de conclure que cette répartition suit une répartition similaire à une répartition de Poisson ($\lambda = 1$)(cf. graphique ci-dessous). En conséquence une modélisation linéaire n'est pas possible, une modélisation à l'aide d'une régression de Poisson est donc proposée.

Graphique : nombre d'accidents par rue de Paris, répartition



Gestion des valeurs manquantes

Des valeurs manquantes pour le nombre de voies et la longueur totale de la rue ont été corrigée en imputant la moyenne de ces variables en fonction du type de voie. L'hypothèses est que le type de voie donne une information importante sur le nombre de voie et la longueur de la rue.

Lien entre les variables

La matrice de corrélation est analysée dont l'extrait les plus significatifs sont reportées ci-dessous. Les accidents à vélos sont en nombre positivement corrélés au nombre d'accidents total (hors accident à vélos). Ce qui implique qu'un nombre restreint de rues concentre la majorité des accidents.

variables	nb_accidents_total	nb_accidents_velo	nb_voies	nb_accidents_horsvelos
nb_accidents_total	1,00	0,74	0,32	0,99
nb_accidents_velo		1,00	0,19	0,65
nb_voies			1,00	0,32
nb_accidents_horsvelos				1,00

Une régression de Poisson simple sur ces deux variables confirme l'intuition que la longueur totale et le nombre de voie sont positivement corrélée à ces variables. En effet, ces deux variables contiennent une information à propos de la densité de circulation. Elles sont donc des variables de contrôle majeurs dans la modélisation qui suit.

Modélisation

Le modèle suivant est :

$$Nb\ accidents\ vélo = \alpha_0 + \alpha_1 \times (Nb\ accidents) + \Pi \times L + \beta_1 \times (Longueur) + \beta_2 \times (Nb\ voies)$$

Où Π est la matrice des coefficients qui se rapportent aux variables *L* de longueur exprimé en proportion de la longueur totale de la rue. Plusieurs compositions pour la matrice de longueur sont proposés ci-dessous et les résultats de la régression y sont résumés :

	Dependent variable:		
	Nombre d'accidents de vélos		
	(1)	(2)	(3)
longueur	0.001***	0.001***	0.001***
	(0.0001)	(0.00005)	(0.0001)
l_total	0.0001***	0.0002***	0.0001***
	(0.00003)	(0.00003)	(0.00003)
l_sens_contresens			0.181
			(0.111)
l_sens_general			0.207*
			(0.111)
l_couloirbus	0.015		-0.185*
	(0.011)		(0.111)
l_sansmarquage	-0.018		-0.132
	(0.062)		(0.108)
l_bandecyclable	0.026**		-0.180
	(0.012)		(0.112)
l_pistecyclable	0.025**		-0.174
	(0.012)		(0.111)
l_continuitecyclable	0.588***		0.517***
	(0.075)		(0.093)
nb_accidents_horsvelos	0.038***	0.038***	0.038***
	(0.004)	(0.004)	(0.004)
nb_voies	0.469***	0.482***	0.473***
	(0.040)	(0.039)	(0.040)
Constant	-3.630***	-3.614***	-3.651***
	(0.093)	(0.089)	(0.094)
Observations	5,842	6,005	5,842
Log Likelihood	-1,684.454	-1,736.807	-1,682.365
Akaike Inf. Crit.	3,388.908	3,483.613	3,388.730
Note:		*p<0.1; **p<0.05; ***p<0.01	

On remarque que minimise la perte d'information est le modèle n°3 pour lequel le critère AIC est le plus faible. Par ailleurs, si l'existence d'un couloir de bus tend à avoir un lien négatif avec le nombre d'accidents à vélos, la continuité cyclable a un lien positif et l'étendue de l'effet (comparable car les unités sont identiques) est plus élevé. On observe donc que les continuités cyclables moins protégés des voitures sont le lieu d'accidents. Ce qui ne peux être conclue pour autre typologie de voie.

5. ANNEXES

Descriptif des variables de la table totale

La table totale est celle résultant de la concaténation des 4 tables fournies par le ministère de l'intérieur (Cf. Partie 3.)

nom de variable	description	modalités (si qualitative)
Num_Acc	Identificateur de l'accident	
mois	mois de l'année	1, 2, …, 12
jour	jour du mois	1,2…,31
lum	Lumière : conditions d'éclairage dans lesquelles l'accident s'est produit	1 – Plein jour 2 – Crépuscule ou aube 3 – Nuit sans éclairage public 4 - Nuit avec éclairage public non allumé 5 – Nuit avec éclairage public allumé
agg	Localisation :	1 – Hors agglomération 2 – En agglomération
int	Intersection	1 – Hors intersection 2 – Intersection en X 3 – Intersection en T 4 – Intersection en Y 5 - Intersection à plus de 4 branches 6 - Giratoire 7 - Place 8 – Passage à niveau 9 – Autre intersection
atm	Conditions atmosphériques	1 – Normale 2 – Pluie légère 3 – Pluie forte 4 – Neige - grêle 5 – Brouillard - fumée 6 – Vent fort - tempête 7 – Temps éblouissant 8 – Temps couvert 9 – Autre

col	Type de collision :	1 – Deux véhicules - frontale 2 – Deux véhicules – par l'arrière 3 – Deux véhicules – par le coté 4 – Trois véhicules et plus – en chaîne 5 – Trois véhicules et plus - collisions multiples 6 – Autre collision 7 – Sans collision
adr	Adresse postale	
circ	Régime de circulation	1 – A sens unique 2 – Bidirectionnelle 3 – A chaussées séparées 4 – Avec voies d'affectation variable
prof	Profil en long décrit la déclivité de la route à l' endroit de l'accident	1 - Plat 2 - Pente 3 - Sommet de côte 4- Bas de côte
plan	Tracé en plan	1 – Partie rectiligne 2 – En courbe à gauche 3 – En courbe à droite 4 – En « S »
larrout	Largeur de la chaussée affectée à la circulation des véhicules ne sont pas compris les bandes d'arrêt d'urgence, les TPC et les places de stationnement	
surf	Etat de la surface	1 - normale 2 - mouillée 3 - flaques 4 - inondée 5 - enneigée 6 - boue 7 - verglacée 8 - corps gras - huile 9 - autre
infra	Aménagement - Infrastructure	1 – Souterrain - tunnel 2 – Pont - autopont

		3 – Bretelle d'échangeur ou de raccordement 4 - Voie ferrée 5 – Carrefour aménagé 6 – Zone piétonne 7 – Zone de péage
man* (quantitative)	Nombre des manœuvres de chaque type précédant l'accident	manDepa = dépassement manTour = tournant manDep = déportation manAutre = autre
nbr*	Nombre de vehicules de chaque type d'un accident	nbrVelo = velo nbrVehiL = poids lourd nbrVehiO = vehicule ordinaire nbrAutres = autres
a_*	Nombre de personne de chaque tranche d'âge	a_60P = >60 ans a_60 = 25 – 60 a_25= 15 – 25 a_15 = 0 - 15
g_*	Nombre de bléssé de chaque catégorie	g_I = Indemne g_H = Hospitalisé g_T = Tué g_B = Bléssé leger
u_*	Nombre d'équipement de protection utilisé	u_c = casques u_r = équipement réfléchissant
temps	periode de la jounée	1 = matin heure de pointe 2 = matin 3 = après midi 4= après midi heure de pointe 5 = soir
jourSem	jour de la semaine	lundi, mardi….