Projet M8 : Cartographie des zones routières dangereuses en fonction de l'âge et du type de conducteur

ALBERT David, SCHAEFFER Marion

Juin 2017

Table des matières

Introduction

Le projet M8 intervient au quatrième semestre de notre formation d'ingénieur et a pour but de nous laisser entièrement autonomes sur la réalisation d'un projet. En effet, nous avons pu mener différents projets durant notre formation mais aucun de cette envergure car c'est la première fois qu'on nous laisse libre sur le choix du sujet. De même, aucune consigne ne nous a été imposée car chacun a dû choisir sa propre problématique et ainsi choisir quels éléments et outils vus en cours utiliser afin de mener à bien ce projet.

Pour notre part, nous avons décidé de travailler sur les accidents de la route. En effet, étant jeunes conducteurs, nous avons été particulièrement sensibilisés lors de l'apprentissage du code de la route et de nos premières heures de conduite, aux dangers présents sur les axes routiers. Nous avons ensuite appris, grâce à un de nos papas gendarme, que lors de chaque accident de la route, chaque force de l'ordre présent sur place se devait de remplir une sorte de grand questionnaire concernant l'accident. En cherchant, nous avons réussi à trouver ses données sur le site de la sécurité routière.

Une fois les données trouvées, il a fallut chercher une problématique afin de savoir sous quel axe les étudier. Dès le début nous avons souhaité cartographier les accidents. L'idée première était de pouvoir faire une carte avec un code couleur en fonction de la dangerosité de la route. Puis nous est venu l'idée de construire un avertisseur sonore pour prévenir le conducteur lors de l'approche d'une zone dangereuse. Nous avons donc décider de définir les zones routières dangereuses en fonction de l'âge et du type de conducteur. Une fois analysées, elles pourront être utilisées afin de les cartographier dans le cadre du projet M8 et pour essayer de construire un petit avertisseur sonore en tant que projet personnel.

Afin de mener à bien ce projet, nous consacrerons le début de ce rapport à la description de nos données. Nous détaillerons ensuite le traitement fait sur les données afin de justifier les interprétations faites dans la partie suivante. Enfin, nous conluerons sur les résultats et ce que nous a apporté ce projet.

Première partie

Présentation des données

Chapitre 1

Téléchargement des données

Comme nous l'avons dit précedemment, nous avons trouvé nos données sur le site de la sécurité routière. Elles se présentaient sous la forme de 4 tableurs excel. Un concernant les lieux d'accidents, un pour les véhicules impliqués, un pour les usagers sur place et un permettant de décrire les caractéristiques de l'accident. Il y avait également un pdf détaillant toutes les notations et la signification de toutes les valeurs entrées dans les tableurs.

Nous avons dans un premier temps téléchargé les données relatives aux années 2014 et 2015. Puis en les parcourant, nous nous sommes rendus compte que, à elle seule, l'année 2015 rendait compte de plus de 60000 accidents et impliquait plus de 130000 usagers. Nous avons donc décidé de nous concentrer uniquement sur l'année 2015 afin de simplifier l'étude et pour qu'elle soit la plus récente possible. Il est également important de souligner que nos variables étaient pratiquement toutes qualitatives (seuls l'âge et les coordonnées de longitude/latitude sont quantitatives). Nous avons donc dû adapter les outils du cours et faire énormément de recherches par nous même pour appliquer les modèles statistiques.

Pour pouvoir travailler chez nous, nous avons décidé d'installer et de travailler sur Octave, donc tout le code que vous trouverez dans la suite du rapport est écrit sur Octave.

Nous avons donc dû commencer par transformer le tableur excel en une matrice Matlab. Voici un exemple avec le fichier usager.

```
%Chargement des données
fichier_usagers = fopen('usagers_2015.txt','r');
%On regarde les types de chaque variable
[usager nb_valeurs_mat_usagers] = fscanf(fichier_usagers, '%lu %u
↪  %u %u %u %u %u %u %u %u %u',[11 Inf]);
```

Chapitre 2

Contenu des données

2.1 Lieux

Dans ce tableur, les informations concernant les lieux des accidents ont été renseignées. On peut ainsi y trouver le numéro de l'accident, le type de route ainsi que le nombre et le type de voie, leur numéro, le profil et le plan de la route, l'état de la surface, les infrastructures présentes, etc. Comme ce sera le cas pour toutes les variables également décrites par la suite, les n différentes modalités prises par la variable sont représentées par des chiffres allant de 1 à n. Ce tableur a engendré une matrice de dimensions :

```
[lieu_nb_individu lieu_nb_variable]= size(lieu); % 58654 X 16
```

2.2 Caractéristiques

La partie caractéristique représente les infos principales liées à l'accident. Le numéro d'accident est donc associé à une date, une luminosité, les conditions météorologiques, une localisation, un type d'intersection et la description de la collision, la lattitude et longitude. Cela nous a fournit une matrice :

```
[car_nb_individu car_nb_variable]= size(car); % 58654 X 14
```

2.3 Véhicules

La partie véhicule ressence les véhicules impliqués dans chaque accident. On retrouve ainsi le numéro d'accident, le type de véhicule, les obstacles heurtés ainsi que la position du choc, le type de manoeuvre en cours et le nombre d'occupants du véhicule. La matrice représentative a des dimensions de :

```
[vehicule_nb_individu vehicule_nb_variable]= size(vehicule); %99778 X 8
```

2.4 Usagers

Ce tableur ressence toutes les personnes mises en cause dans l'accident. Il comprend donc les numéros d'accident et des véhicules impliqués, la place de l'occupant dans ce dernier, la gravité de ses blessures, le sexe, l'année de naissane, le type de trajet, les mesures de sécurité mises en place ainsi que la présence d'un piéton. Cela représente une matrice de :

```
[usagers_nb_individu usagers_nb_variable]= size(usagers); %130378 X 11
```

Chapitre 3

Utilisation des données

Tout au long de ce projet, nous avons essayé d'effectuer les différents tests sur l'ensemble des données afin de pouvoir toutes les traiter et chercher au maximum les relations qu'il existait entre elles, et cela sans à prioris humains. Cependant, nous nous sommes vite rendu compte que cela serait difficile. En premier lieu, les matrices n'avaient pas les mêmes dimensions : il y a bien le même nombre de lieu d'accidents que de caractéristiques d'accidents mais il n'y a pas le même nombre d'usagers et de véhicules impliqués. Ainsi le seul lien pour comparer des variables des différentes matrices n'était autre que le numéro d'identification de l'accident (répertorié dans la première colonne à chaque fois). Nous avons, tout de même, réussi à confronter les données de plusieurs matrices mais avec de plus grandes connaissances en gestion de bases de données (utilisation de Système adaptée tel que MySQL par exemple) nous aurions probablement pû effectuer de nombreuses requêtes qui aurait sûrement été moins coûteuses en temps de calcul.

De plus, quand nous avons voulu faire notre régression, cela a également été très fastidieux car nous ne possédions que de très peu de variables quantitatives : les années de naissance, les dates et horaires d'accidents ainsi que les adresses et coordonnées GPS. Les seuls critères potentiellement intéressants à comparer étaient donc les années de naissance et les coordonnées GPS (plus facile à traiter que les adresses) mais n'étant pas dans les mêmes matrices, le problème précédemment cité s'est posé.
Pour l'ACP et les tests du Chi2, nous nous sommes résolus à comparer les variables appartenant aux mêmes matrices. Au final, les deux matrices que nous avons traitées en priorité, par leur intérêt accru sont la matrice des caractéristiques des accidents et celle des caractéristiques des usagers.

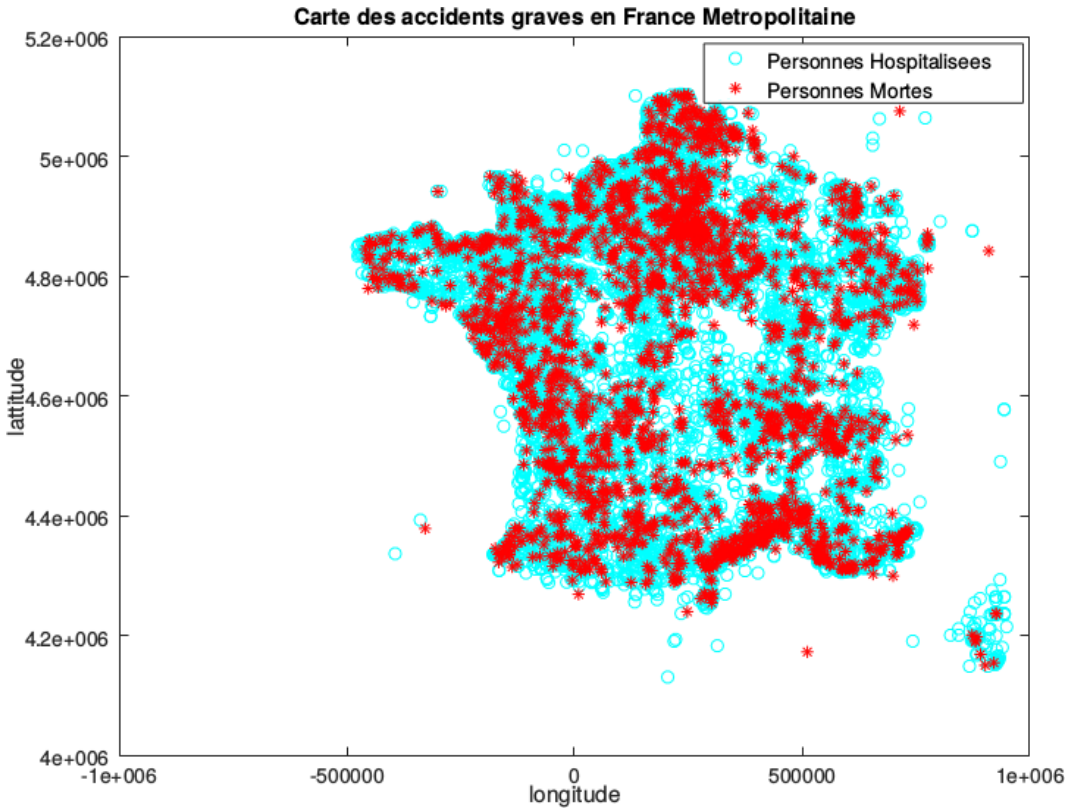
Deuxième partie

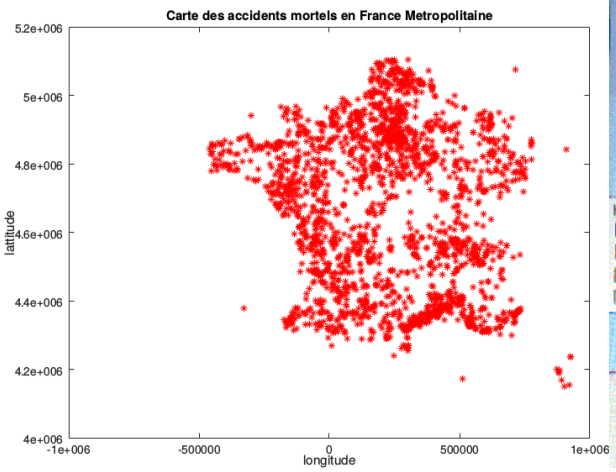
Traitements des données

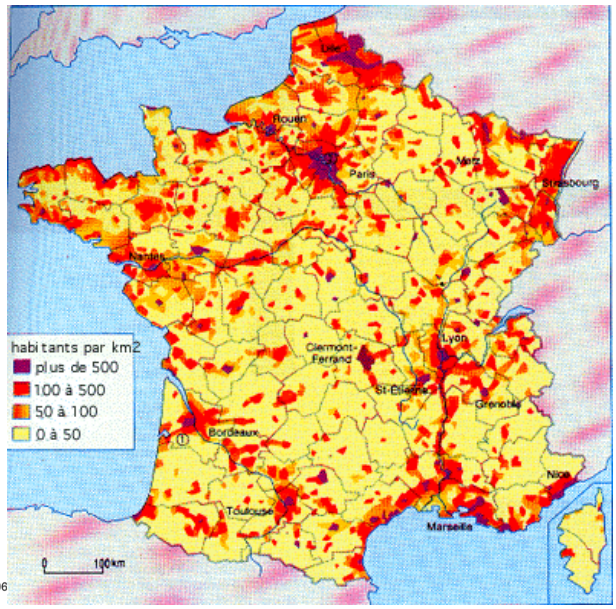
Chapitre 4

Fréquences

Analyse brute de carte







Nous avons ici afficher le nuage de point des coordonnées GPS (latitude/longitude) des accidents très graves en France. En étudiant ces graphiques, nous constatons que les zones qui semblent être propices aux accidents graves sont Paris et ses alentours, la côte méditerranéenne, le Nord de la France et la région de la Loire. Par comparaison avec la carte démographique de France, nous pouvons facilement avancé l'hypothèse que les zones les plus peuplées sont propices aux accidents, tandis que les zones vide (moins de 50 $habitants/km^2$) ne présentent que très rarement des accidents graves. Cette étude cartographique ne peut donc pas être concluente et il serait préférable de faire une analyse de fréquences précise.

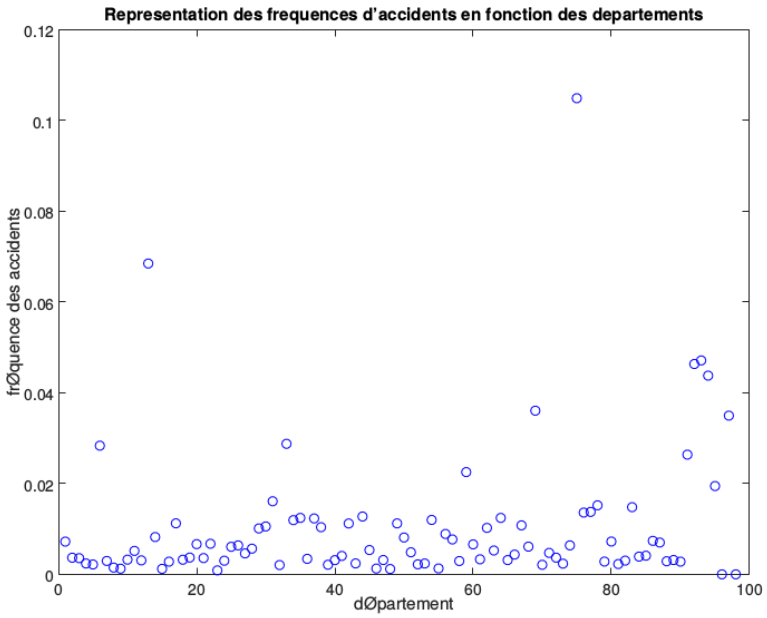
Analyse par département

Pour commencer nos tests, nous avons voulu calculer les fréquences des accidents par départements. Nous avons commencé par calculer le nombre d'accidents par département.

```
for i=1 : n
        if ((car(i,end)==201) | (car(i,end)==202)) %Pour la Corse
                X(20)= X(20) + 1;
        else
                X(floor(car(i,end)/10)) = X(floor(car(i,end)/10))
↪  +1; %Les départements sont représentés avec un 0 en plus à la
↪  fin
   endif
endfor
```

Nous avons ensuite calculé les fréquences, vérifié que leur somme valait 1, nous les avons également tracé.

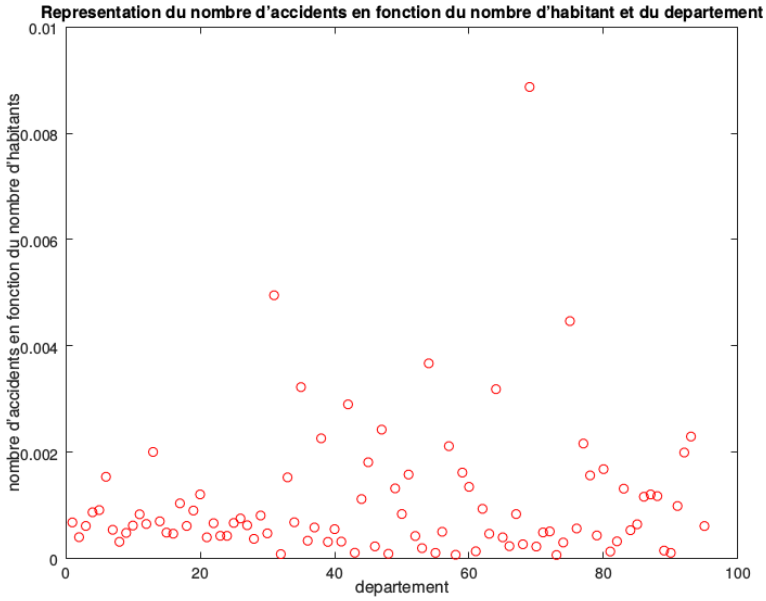
```
f = X/n;
sum(f)
```



On remarque que certains départements sortent du lot, notamment les Bouches du Rhone (point en haut à gauche) et Paris (point en haut à droite). Il y a donc bien plus d'accidents dans ces départements que dans d'autres. Pourtant, ce résultat est à nuancer, en particuler avec le nombre d'habitants car ces 2 départements font partit des plus peuplés. Nous avons donc décidé de diviser le nombre d'accidents par départements par le nombres d'habitants du département. Initialement nous voulions trouver le nombre de personnes conduisant dans chaque départements mais ces valeurs se sont révélés impossible à trouver. Nous avons donc garder l'idée du nombre d'habitants.

```
y=X./pop %pop est un vecteur contenant le nombre d'habitants par
↪  département
```

Nous avons ainsi obtenu le graphique suivant :



Finalement, on observe des différences avec le graphique précedement tracé. Les Bouches du Rhône et Paris continuent à sortir du lot mais c'est le département du Rhône (le plus haut à droite) qui domine largement ainsi que l'Hérault (le plus haut à gauche). D'autres département tendent à se démarquer comme on peut le voir dans la tranche des valeurs entre 0.003 et 0.006 sur l'axe des ordonnées.
Il est donc important de prendre en compte le nombre d'habitants par département pour juger les zones dangereuses car sinon les chiffres ne veulent rien dire. Ils nous permettent ainsi de savoir réellement quelles sont les zones de France où il y a le plus de chance d'avoir un accident (sans tenir compte de la gravité).

Chapitre 5

Analyse en Composantes Multiples

Comme nous disposons de variables qualitatives pour la grande majorité, nous avons dû faire une ACM et non une ACP. Ce travail nous a demandé beaucoup d'investissement pour mener des recherches bibliographiques sur les méthodes d'application et d'interprétation de l'ACM.
Nous avons commencé par créer les tableaux disjonctifs complet propre à chacune de nos variables. Le tableau disjonctif complet est un tableau disjonctif tel que nous l'avons vu pour la distance du Chi2 mais au lieu de ne traiter que les modalités d'une variable qualitative à la fois, on rassemble les modalités de toutes les variables à traiter à la fois. Par exemple ici pour la matrice des caractéristiques :

```
x = [car(:,6) car(:,7) car(:,9) car(:,10)];
TDC_luminosite = [x(:,1)==1 x(:,1)==2 x(:,1)==3 x(:,1)==4
↪  x(:,1)==5];
TDC_agglo = [x(:,2)==1 x(:,2)==2];
TDC_temps = [x(:,3)==1 x(:,3)==2 x(:,3)==3 x(:,3)==4 x(:,3)==5
↪  x(:,3)==6 x(:,3)==7 x(:,3)==8 x(:,3)==9];
TDC_collision = [x(:,4)==1 x(:,4)==2 x(:,4)==3 x(:,4)==4
↪  x(:,4)==5 x(:,4)==6 x(:,4)==7];
TDC_car = [TDC_luminosite TDC_agglo TDC_temps TDC_collision];
```

Nous avons multiplié les tableaux disjonctifs transposés par eux-mêmes pour obtenir une table de Burt, qui est donc une sorte de table de contingence généralisé.

```
B = TDC_car'*TDC_car;
```

Nous avons ensuite tenté de calculer les profils lignes et profils colonnes normalisés à partir des effectifs marginaux.

```
[ligne colonne] = size(B)
effect_lignes = sum(B, 2)
effect_colonnes = sum(B,1)
prof_lignes_norm = B./sqrt(effect_lignes*ones(1, colonne))
prof_colonnes_norm = B./sqrt(ones(ligne, 1)*effect_colonnes)
% On vérifie l'égalité du résultat
prof_colonnes_norm - prof_lignes_norm'
```

Tout ce travail a été fait dans le but de tenir compte du nombre de modalités possibles de chacune des variables. Nous avons ensuite réalisé une ACP comme vue en cours sur les profils colonnes (le calcul peut aussi bien se faire sur les profils lignes, le résultat est le même). On a donc pu afficher le nuage de point représentant les individus (les accidents) et celui représentant les modalités. Pour les modalités proches - par exemple les modalités liées au mauvais temps (pluie, grèle, neige, etc) sont proches dans le nuage de points - nous avons préféré afficher uniquement le barycentre des modalités pour plus de simplicité et de compréhension dans l'analyse (« mauvais temps »).

```
[vect_p val_p] = eig(prof_colonnes'*prof_colonnes);
[n p] = size(prof_colonnes'*prof_colonnes);
V2 =  [vect_p(:,end) vect_p(:,end-1)]
U = TDC_car*V2;
Vn = vect_p*sqrt(val_p)/sqrt(n)
```

En réalisant l'ACP, nous nous sommes rendus compte que seules les matrices lieux et caractéristiques étaient facilement expoitables pour ces tests car il manquait plusieurs valeurs dans ces matrices que nous avons remplacé par des '0' pour pouvoir charger le fichier. Seulement en divisant par les effectifs marginaux (donc par 0), des NaN sont apparus et nous ne pouvions donc pas calculer les valeurs propres sur de tels résultats.

Nous avons pu représenté le pourcentage d'informations, les individus, les variables mais aussi certaines caractéristiques spécifiques pour la matrice caractéristique et nous les commenterons dans la partie Interprétation plus loin dans le dossier.

Chapitre 6

Régression linéaire

Nous avons choisis de faire une régression linéaire sur les variables âges, latitude et longitude. La variable à expliquer sera l'âge et les variables explicatives seront la latitude puis la longitude.
Nous avons commencé par faire en sorte d'obtenir des matrices de même taille. Nous avons donc associé chaque latitude à l'âge auquel il correspond.

```
age = 2017.-usager(:,11)';
latitude = car(:,end-2);
nb_personnes = length(age);
X = [lattitude(1)] ;
j = 1;
for i=2:nb_personnes
        if (usager(i,1)==car(j,1))
                X = [X latitude(j)];  %On reste sur la même
↪  latitude si le numéro d'accident est toujour le même
        else
                j = j+1; %On passe à la latitude suivante si les
↪  numéros d''accidents sont différents
                X = [X latitude(j)];
        endif
endfor;
```

Nous avons ensuite enlever les valeurs aberrantes, c'est à dire les valeurs manquantes qui étaient écrites comme des 0 et les valeurs des laltitudes correspondantes DOM-TOM.

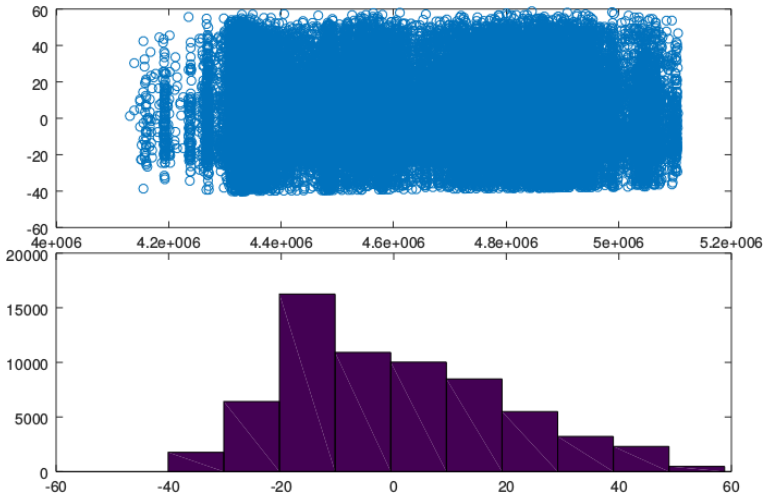
```
non_abberant = find((age<100) & (X!=0) & (X>4000000)); %
↪  recherche des indices non aberrants
age = age(non_abberant);
X = X(non_abberant);
```

```
Nous avons ensuite tracé les latitudes des accidents en fonction
↪ de
l'âge des individus:

Nous avons calculé les paramètres de la régression.

\begin{minted}{octave}
x = [X', ones(n,1)];
a = (x'*x)\(x'*age') %a = -3.5319e-006 et b = 5.7317e+001
```

Voici un aperçu des résidus. On remarque qu'ils suivent bien une structure gaussienne.



Nous avons ensuite cherché à faire un test d'indépendance. La p_{valeur} est largement supérieure à 0.05 donc les variables sont indépendantes.

```
s2 = e'*e/(n-2);
m_latitude = mean(X);
s2_latitude = std(X)^2;
%H0 : On suppose l'indépendance des valeurs
%H1 : On suppose la dépendance des valeurs
T = a(1)/sqrt(s2/s2_latitude);
ddl = n-2;
pval = 2*(1-tcdf(abs(T), ddl)) %0.96596
```

Il a fallut finir par évaluer la qualité de la régression au sens des moindres carrés :

```
SCT = sum((age-mean(age)).^2);
SCM = sum((x*a-mean(age)).^2);
```

```
SCR = e'*e;
R2 = 1 - SCR/SCT %=0.0018179
```

Le R2 est très proche de 0, la régression est donc très mauvaise. Le modèle n'est pas bon, les variables sont indépendantes.
Nous avons refait cette étude avec la longitude, nous ne détaillerons pas le code car il est similaire mais voici les résultats obtenus :

— $a = -8,0716 \times 10^{-7}$
— $b = 41,077$
— $p_{valeur} = 0.99041$
— $R^2 = 1.4456 \times 10^{-4}$

Troisième partie

Interprétation

Chapitre 7

Test du χ_2

Procédure à suivre

De la même manière que nous avons effectué l'ACP sur les matrices caractéristiques et lieux, nous avons effectué le test du Chi2 sur les matrices usagers et caractéristiques afin de réaliser un traitement intéressant sur un maximum de données.

Pour tous les tests, nous avons donc chercher à répondre à la question : « Les variables sont-elles dépendantes ? ».

Ainsi nous avons posé l'hypothèse H0, dans laquelle nous supposons l'indépendance des variables et H1, où nous suppossons la dépendance des variables. Pour la matrice usagers, nous avons tout d'abord testé la dépendance entre chaque variable de la matrice en fonction de la gravité, comme ci-dessous :

```
for i=1:8
        switch i
                case {1}
                        var_TDC = TDC_place_vehicule;
            case {2}
                        var_TDC = TDC_type_usager;
                case {3}
                        var_TDC = TDC_sexe;
                case {4}
                        var_TDC = TDC_type_trajet;
                case {5}
                        var_TDC = TDC_mesure_secu;
                case {6}
                        var_TDC = TDC_loca_pieton;
                case {7}
                        var_TDC = TDC_action_pieton;
                case {8}
                        var_TDC = TDC_compagnie;
        endswitch
```

```
        O=var_TDC'*TDC_gravite
        m=sum(O);
        n=sum(O,2);
        eff = sum(n);
        T=n*m/eff;
        Dchi2 = sum(sum((O-T).^2./T))
        ddl = (length(m)-1)*(length(n)-1);
        pval = 1-chi2cdf(Dchi2, ddl)
endfor
```

En considérant tous les usagers nous avons trouvé que toutes les p-valeurs sont égales à 0. Deux conclusions peuvent donc être tirées : soit toutes les variables ont une très forte dépendance avec la gravité, soit le test utilisé n'est pas adapté. Il semble évidemment préférable dans notre cas de considéré que le test n'est pas adapté. En effet, il se pourrait que le grand nombre d'usagers (130 378) et la rareté d'apparition de certaines modalités de variables puissent être à l'origine d'une très grande distance du Chi2 et donc de la p_{valeur} nulle.

Pour comblé ce problème nous avons donc décidé de restreindre l'échantillon d'étude pour le calcul de la p_{valeur}. Nous avons donc considéré que nous nous restreindrions à des échantillons de 200 individus pour pouvoir avoir des p_{valeur} mesurables et comparables. Ainsi, pour savoir si des variables sont indépendantes ou non, nous les comparons entre elles et nous les comparons au risque de première espèce α que l'on déterminera (pour 200 individus, $\alpha = 0.05$ convient).

Enfin, après réflexion, nous nous avons pensé que prendre un seul échantillon de 200 individus ne pouvait pas être représentatif de l'ensemble des individus (des dizaines de milliers) et que cela risquait fortement de générer des résultats fossés par un échantillon particulier. Donc nous avons décidé de prendre N échantillons de 200 individus et de calculer les N p_{valeur} associées (ici N=1000). La moyenne de ces p_{valeur} nous semble plus représentative de la p_{valeur} du couple de variables pour l'ensemble des données.

Le code qui va bien est le suivant :

```
% Pour les tests de Chi2 on utilisera :

taille_echantillon = 200; % nb d'individus
nb_de_pvaleurs = 1000; % N échantillons/p-valeurs

%Ho : on suppose l'indépendance entre la variable 1 et la
↪  variable 2 (pour un échantillon de 200 personnes pris
↪  aléatoirement dans la base de données)
%H1 : on suppose la dependance

p_vals = [];
```

```
for i = 1:nb_de_pvaleurs
        debut_echantillon = floor(130000*rand(1,1)); % On prend
↪  une valeur aléatoire de départ
        O=TDC_var1(debut_echantillon:debut_echantillon+taille_echantillon,:)'*
↪  TDC_var2(debut_echantillon:debut_echantillon+taille_echantillon,:);
        m=sum(O);
        n=sum(O,2);
        T=n*m/sum(n);
        Dchi2 = sum(sum((O-T).^2./T));
        ddl = (length(m)-1)*(length(n)-1);
        pval = 1-chi2cdf(Dchi2, ddl);
        if not(isnan(pval))
                p_vals = [pval ; p_vals];
        endif
endfor
pval_m = mean(p_vals)
% On décide de la dépendance ou non des variables 1 et 2 en
↪  fonction de la moyenne des p-valeurs
```

Résultats

Nous avons donc testé plusieurs couples de variables de la matrice usagers :

— sexe / gravité : $p_{valeur} = 0.35363$
Interprétation : $p_{valeur} > 0.05$ donc on considère qu'il y a indépendance entre le sexe de l'usager et l'état de gravité de l'accidenté.

— type de trajet / gravité : $p_{valeur} = 0.27582$
Interprétation : $p_{valeur} > 0.05$ donc on considère qu'il y a indépendance entre le type de trajet effectué par l'usager et l'état de gravité de l'accidenté.

— type d'usager / gravité : $p_{valeur} = 0.02395$
Interprétation : $p_{valeur} < 0.05$ donc on considère qu'il y a dépendance entre le type d'usager (conducteur, passager, piéton, etc) et l'état de gravité de l'accidenté.

— place dans le véhicule / sexe : $p_{valeur} = 0.047489$
Interprétation : Malgré ce que l'on pourrait penser, la $p_{valeur} < 0.05$ donc on considère qu'il y a dépendance entre la place occupé par l'usager et son sexe. Cependant la valeur étant proche du α , il faudrait réaliser des tests supplémentaires.

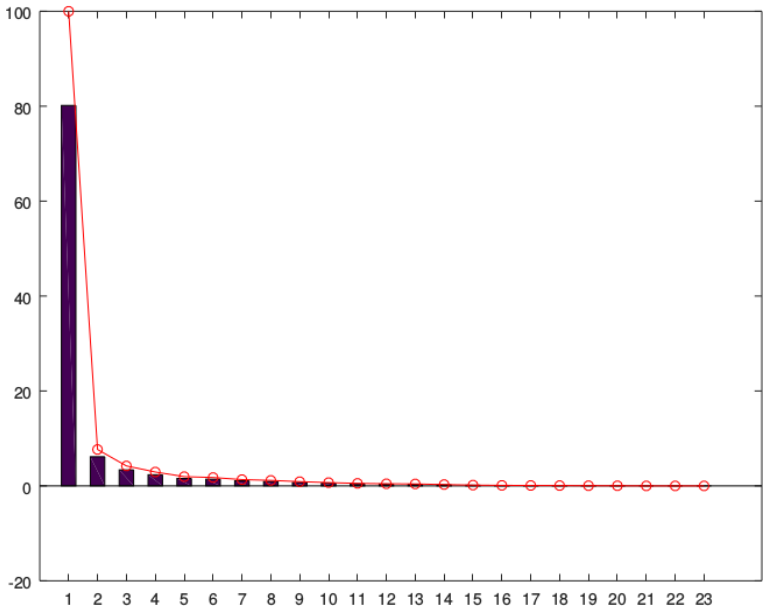
— luminosité / temps : $p_{valeur} = 0.10267$
Interprétation : Malgré le nuage de point de l'ACM (voir interprétation de l'ACM), on ne peut pas affirmer qu'il y a dépendance entre la luminosité et le temps.

Chapitre 8

Axes principaux

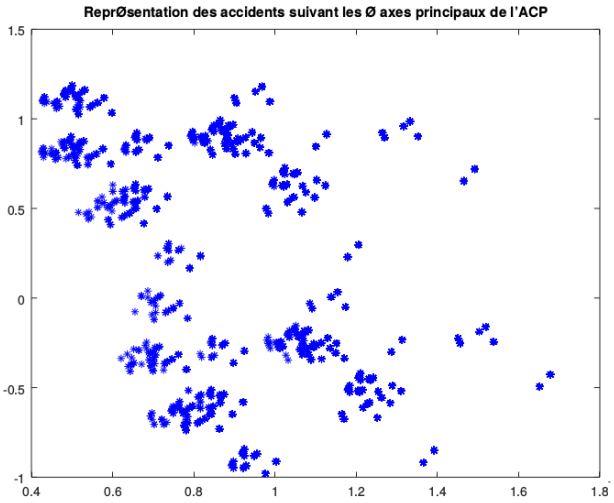
Nous allons ici interpréter les résultats liés à l'Analyse en Composantes Multiples sur la matrice des Caractéristiques de l'accident.

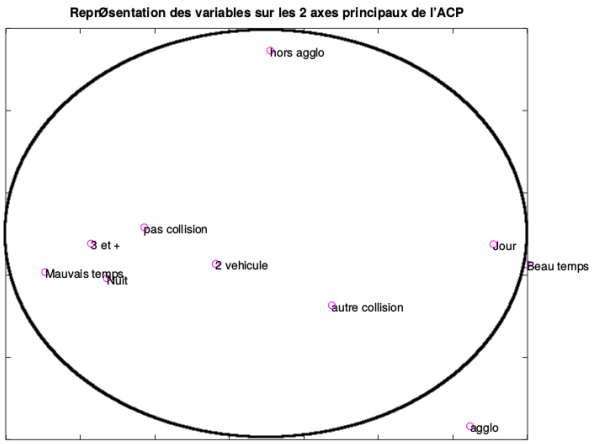
Pourcentage d'information



Plus de 85% de l'information est représenté sur les deux vecteurs propres principaux (ceux liés aux valeurs propres les plus grandes). Respectivement, c'est 80% de l'information qui est représenté sur le premier axe et 5 à 7% sur le second.

Affichage des individus et des variables





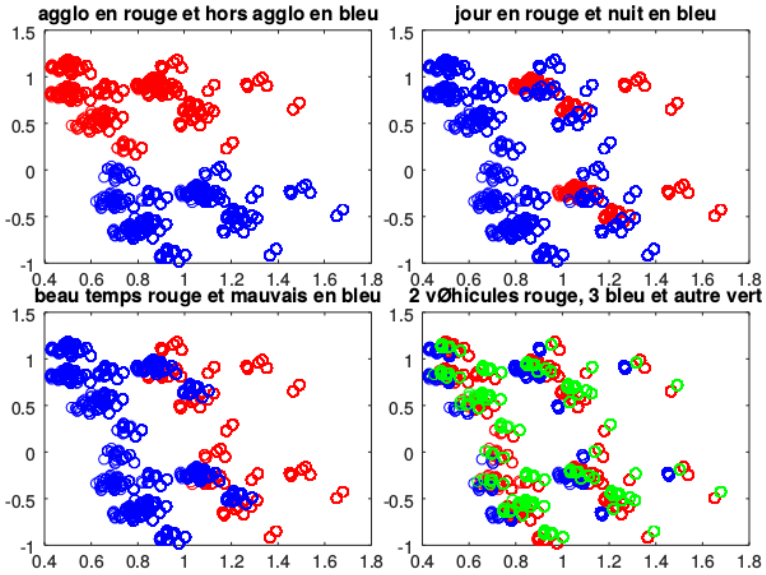
Nous pouvons voir que les individus forment globalement 4 gros groupes (globalement à droite) et 2 petits groupes avec une séparation très marquée Nord-Sud et une séparation plus futile Ouest-Est. Quant aux variables, l' « Agglomération » s'oppose au « Hors Agglomération » (elle correspond aux individus qui s'oppose de façon Nord - Sud), le « Beau Temps » s'oppose de même au « Mauvais Temps » et le « Jour » s'oppose à la « Nuit ». De plus « Beau Temp » et « Jour » ont une distance faible et de même pour le « Mauvais Temps » et la « Nuit ». Ces variables étant proches entre elles et proches du cercle trigonométrique, elles représentent des groupes d'accidents typiques qui sont les accidents la nuit par mauvais temps et les accidents le jour par beau temps.

Les variables recenssant les types de collisions étant centrées, elles ne peuvent pas caractériser fortement un type d'accident.

Pour finir sur l'analyse des nuages de points, on peut globalement dire qu'il semble qu'il y a moins d'accidents le jour et par beau temps que la nuit et par mauvais temps car la densité de points est plus forte à gauche.

Vérifications

Nous avons représenté de couleurs différentes les accidents en fonction des modalités pour appuyer l'interprétation de l'ACM.


agglo en rouge et hors agglo en bleu
jour en rouge et nuit en bleu
beau temps rouge et mauvais en bleu
2 vØhicules rouge, 3 bleu et autre vert

Chapitre 9

Linéarité

Lorsque nous avons tracé les individus, nous nous sommes doutés qu'il n'existait pas de relation linéaire. De même, on se doute bien avant toute étude qu'il n'y en a pas. En effet il semblerait compliqué de relier le lieu de l'accident à l'âge de la victime. Pourtant, nous avons dès le début voulu faire nos tests sans aucun préjugé humain. Nous avons donc décidé de tout de même réaliser la régression.
Dès le calcul des paramètres, nous pouvions porter un jugement sur la qualité de la régression. En effet, dans les 2 cas le coefficient directeur a s'approchait de 0 et l'ordonnée à l'origine b d'une valeur moyenne.
Dans les 2 cas les résidus étaient répartis de manière aléatoire, on pouvait très clairement voir une forme gaussiène sur l'histogramme, donc du point de vue des résidus le modèle est bon.
Nous avons décidé d'effectuer un test de student pour vérifier l'indépendance des variables. L'hypothèse H0 était l'indépendance et l'hypothèse H1 la dépendance. Dans les 2 cas, les p_{valeur} était très largement supérieur à 0.05 (elles étaient même proche de 1), ce qui nous a permis de rejetté H0. Nous aurions également pu, comme dans la partie sur le Chi2, traité un échantillon de valeur et faire leur moyenne pour s'assurer de la justesse des résultats.
Enfin, un dernier test nous a assurer de la cohérence de la régression, c'est le calcul du R2. Dans les 2 cas nous avons trouvé des R2 très proche de 0, ce qui nous a confirmé que le modèle posé n'est pas le bon et prouve définitivement l'indépendance des variables.
Avec la régression linéaire, les problèmes des grands jeux de donnés sont normalement mieux traités que par exemple pour le test du Chi2 précedement cité. Nous pouvons donc faire confiance aux résultats trouvés.
Nous aurions aimé pouvoir faire une régression sur des variables plus intéressantes mais elles se sont révélées êtres qualitiatives et nous n'avons pas disposés d'assez de temps pour rechercher les méthodes de calcul et les mettre en pratique.

Conclusion

En conclusion, ce projet nous a appris à manipuler de grosses bases de données, en organisant notre code pour ne pas nous mélanger entre les variables et pouvoir facilement retrouver les éléments dont nous avions besoin. De plus, nous avons appris à nous documenter sur les fonctionnalités propres au logiciel Octave. Concernant les résultats statistiques, nous n'avons pas réussi à obtenir tous les résultats espérés (difficulté lors des tests d'indépendance) mais nous avons tout de même réussi à mettre en place des moyens - qui nous semblent les plus appropriés - pour comparer les données de manière simple et modulable.
Durant notre étude, nous avons donc mis en avant plusieurs faits, tout d'abord nous avons montrer qu'en proportion le nombre d'accidents ne dépend pas forcément du nombre d'habitants. Ensuite nous avons montrer que la gravité des dégâts subit par une personne victime d'un accident dépend très fortement de son rôle dans l'accident (piéton, conducteur, passager). Nous avons aussi pu remarquer une dépendance entre le sexe de la personne et la position qu'elle occupe dans le véhicule.
Nous aurions aussi voulu, par manque de temps nous n'avons pas pu, faire d'autres analyses. Par exemple, nous aurions préféré faire un tableau qui récapitule toutes les dépendances entre variables. Nous aurions aussi aimé faire une régression linéaire entre le nombre d'accident par département et le nombre d'habitants par département. Et étant donné les vastes possibilités que nous offrait les données nous aurions pu imaginer de nombreux autres tests.