

# **Projet Data Mining**

Clustering de données issues des accidents corporels de la circulation routière

**Autrices :**

AGÜERA SANCHEZ Martina (p2509103)

SOUTOU Tamazgha (p1813609)

Master 2 Informatique - Parcours Data Science

2025 - 2026

## Table des matières

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Présentation des données</b>	<b>4</b>
<b>3</b>	<b>Prétraitements des données</b>	<b>6</b>
3.1	Table Lieux . . . . .	6
3.2	Table Usagers . . . . .	7
3.3	Table Véhicules . . . . .	7
3.4	Table Caractéristiques . . . . .	8
3.5	Fusion des données . . . . .	8
3.6	Nettoyage des données . . . . .	10
3.6.1	Création variable gravité moyenne . . . . .	12
3.7	Création de la variable <i>gravité moyenne</i> . . . . .	12
<b>4</b>	<b>Clustering</b>	<b>12</b>
4.1	Analyse et Interprétation du Clustering K-Means Adapté	12
4.1.1	Choix de la méthode . . . . .	12
4.1.2	Modélisation du K-Means adapté . . . . .	13
4.1.3	Évaluation de la qualité des clusters et analyse des paramètres du K-Means adapté . . . . .	13
4.1.4	Analyse des variables continues . . . . .	14
4.1.5	Analyse des variables catégorielles dominantes . .	15
4.1.6	Interprétation générale . . . . .	15
4.1.7	Analyse détaillée des clusters . . . . .	15
4.1.8	Visualisation (PCA) . . . . .	17
4.1.9	Réduction de dimension et visualisation . . . . .	17
4.1.10	Analyse de la distribution et de la répartition des clusters . . . . .	18
4.2	Méthodes spécialisées : K-Modes . . . . .	19
4.2.1	Choix du nombre de clusters et de méthode . . .	20
4.2.2	Distribution des différentes catégories par cluster	21
4.2.3	Description des clusters . . . . .	23
4.3	Méthodes spécialisées : K-Prototypes . . . . .	24
4.3.1	Choix du nombre de clusters . . . . .	24
4.3.2	Distribution des différentes catégories . . . . .	25
4.3.3	Description des clusters . . . . .	27
<b>5</b>	<b>Analyse des résultats</b>	<b>28</b>
<b>6</b>	<b>Conclusion</b>	<b>29</b>

## 1 Introduction

Le but de ce projet est de réaliser un *clustering* sur des données réelles et actuelles, représentatives des problématiques que nous pourrions rencontrer plus tard dans un contexte professionnel. Pour cela, nous avons choisi de travailler sur des données catégorielles, ce qui nous empêchait d'utiliser directement les méthodes classiques vues en cours (K-Means, DBSCAN, etc.) sans les adapter. Ce type de données est néanmoins très réaliste, car les entreprises collectent fréquemment des informations issues d'enquêtes ou de questionnaires, générant ainsi des données catégorielles qu'il faut ensuite analyser.

Pour nous organiser dans ce projet, nous avons établi un planning détaillé avec des tâches distinctes afin d'assurer une répartition équitable du travail. Toutefois, les différentes étapes étant interdépendantes et nos méthodes de travail parfois différentes, nous avons fixé des dates limites pour chaque bloc de tâches afin d'éviter tout retard.

Le tableau suivant présente la planification du projet, incluant la répartition des tâches entre les deux membres du binôme (**Tamazgha** et **Martina**), les échéances principales et les livrables associés. Ce planning est structuré selon les grandes phases du projet : préparation, traitement des données, analyse et synthèse des résultats.

	<b>Tamazgha</b>	<b>Martina</b>	<b>Date</b>
Importation et analyse préliminaires des données	Table lieux et caractéristiques	Table usagers et véhicules	Semaine 1
Prétraitements table par table	Lieux et Caract. : exploration des données, détection de colonnes importantes, vérification des valeurs manquantes	Usagers et Véhicules : suppression colonnes vides, renommage de valeurs, rééquilibrage d'attributs, création variable 'gravité_moyenne'	Semaine 1
Fusion tables	Fusion Lieux et Caractéristiques.	Fusion Usager et Vehicule + fusion finale	Semaine 2
Nettoyage	Nettoyage, sélection des colonnes utiles, identification et remplacement des valeurs manquantes, nettoyage des coordonnées géographiques.	Détection valeurs manquantes. Analyse corrélation entre attributs et suppression d'attributs trop corrélés.	Semaine 2
Transformations	One Hot Encoding et Normalisation des variables numériques	Création table uniquement catégorielle	Semaine 3
Clustering	K-Means adapté : modélisation, visualisation, évaluation, analyse	Recherche de méthodes. KModes et KPrototypes : analyse coût d'inertie	Semaine 3
Analyse résultats	K-Means adapté : Analyse et pertinence sur les données mixtes	KModes et KPrototypes : distribution catégorielle, score de silhouette	Semaine 4
Rapport	Prétraitements sur Lieux et Caractéristiques, fusion, analyse et Interprétation du Clustering K-Means Adapté	Introduction, présentation des données, prétraitements Usagers et Véhicules, clustering et analyse spécialisés, conclusion	Semaine 4

TABLE 1 – Répartition des tâches

## 2 Présentation des données

Pour ce projet, nous avons choisi comme source de données la base des accidents corporels de la circulation routière en France pour l'année 2023, issue du site [data.gouv.fr](http://data.gouv.fr).

Pour chaque accident corporel (c'est-à-dire un accident survenu sur une voie ouverte à la circulation publique, impliquant au moins un véhicule et ayant causé au moins une victime nécessitant des soins) une saisie d'informations est effectuée par les forces de l'ordre (police, gendarmerie, etc.) intervenues sur les lieux. Ces informations sont regroupées dans

une fiche intitulée *Bulletin d'analyse des accidents corporels*. L'ensemble de ces fiches constitue le fichier national des accidents corporels de la circulation, dit *Fichier BAAC*, administré par l'Observatoire national interministériel de la sécurité routière (ONISR).

La base de données extraite du fichier BAAC répertorie l'intégralité des accidents corporels de la circulation survenus en 2023, en France métropolitaine, dans les départements d'outre-mer (Guadeloupe, Guyane, Martinique, La Réunion et Mayotte), ainsi que dans les territoires d'outre-mer (Saint-Pierre-et-Miquelon, Saint-Barthélemy, Saint-Martin, Wallis-et-Futuna, Polynésie française et Nouvelle-Calédonie). Elle fournit une description simplifiée comportant à la fois des informations sur la localisation de l'accident et sur ses caractéristiques (lieu, conditions, véhicules impliqués, victimes, etc.).

Certaines données spécifiques relatives aux usagers ou aux véhicules ont toutefois été occultées afin de préserver la vie privée des personnes physiques potentiellement identifiables, ou lorsque leur divulgation aurait pu révéler des comportements individuels sensibles.

Les usagers en fuite ont été conservés dans la base, ce qui induit des valeurs manquantes pour certaines variables telles que le sexe, l'âge ou encore la gravité des blessures (indemne, blessé léger, blessé hospitalisé).

Un accident corporel (mortel ou non) de la circulation routière, tel que recensé par les forces de l'ordre, répond aux critères suivants :

- il implique au moins une victime ;
- il survient sur une voie publique ou privée ouverte à la circulation ;
- il implique au moins un véhicule.

Chaque accident corporel implique un ou plusieurs usagers, parmi lesquels on distingue :

- **les personnes indemnes** : impliquées dans l'accident, non décédées et ne nécessitant aucun soin médical ;
- **les victimes** : impliquées non indemnes, subdivisées en :
  - **personnes tuées** : décédées sur le coup ou dans les trente jours suivant l'accident ;
  - **personnes blessées** : victimes non tuées, comprenant :
    - **blessés hospitalisés** : victimes hospitalisées plus de 24 heures ;
    - **blessés légers** : victimes ayant nécessité des soins mais non hospitalisées plus de 24 heures.

La base est composée de quatre fichiers (Caractéristiques, Lieux, Véhicules, Usagers) au format CSV. Chaque ligne correspond à un accident identifié par une clé primaire (`Num_Acc`), ou dans le cas de la table

des usagers, à un couple (`Num_Acc`, `Num_Usager`). Chaque table contient une dizaine d'attributs décrivant l'accident. Hormis l'âge de l'usager, la vitesse maximale autorisée, la latitude et la longitude, tous les attributs sont catégoriels. Par exemple, pour la luminosité au moment de l'accident, les forces de l'ordre doivent choisir parmi les modalités suivantes : plein jour, crépuscule ou aube, nuit sans éclairage, nuit avec éclairage allumé ou éteint.

Le jeu de données comprend 125 789 usagers et 93 585 accidents. Il est disponible à l'adresse suivante :

<https://www.data.gouv.fr/datasets/bases-de-donnees-annuelles-des-accidents-cor>

La description complète de tous les attributs des différentes tables est fournie dans le dossier joint.

## 3 Prétraitements des données

### 3.1 Table Lieux

La table Lieux contient 70 860 lignes et 18 colonnes décrivant les caractéristiques géographiques et structurelles des lieux d'accidents. Les principales variables incluent :

- `Num_Acc` : identifiant unique de l'accident.
- `catr` : catégorie de la route.
- `surf` : état de la surface.
- `vma` : vitesse maximale autorisée.

D'autres variables fournissent des informations complémentaires sur la voie : nombre de voies, profil, tracé, type d'infrastructure, situation, etc.

#### Problèmes détectés et solutions apportées

Valeurs manquantes : certaines colonnes présentent un grand nombre de valeurs manquantes : `lartpc`, `v2`, `voie`.

Solution : ces colonnes ont été exclues temporairement des analyses.

Données codées à -1 : de nombreuses variables utilisent la valeur `-1` pour représenter des données manquantes (ex. : `circ`, `prof`, `surf`, `infra`, `situ`, etc.).

Solution : remplacement des valeurs `-1` par `NaN` afin de faciliter l'analyse statistique et la détection d'anomalies.

Types de données incorrects : les colonnes `nbv`, `pr` et `pr1` sont stockées sous le type `object`, alors qu'elles devraient être numériques.

Solution : conversion de ces colonnes en `float` ou `int` selon le cas, après traitement des valeurs manquantes.

Valeurs incohérentes ou non renseignées dans certaines variables : la variable `vma` contient quelques valeurs à -1 (non renseignées) et une forte concentration autour de **50 km/h**, correspondant aux zones urbaines. Solution : remplacement des valeurs -1 par des valeurs manquantes, puis vérification de la cohérence des vitesses selon la catégorie de route.

#### Analyse descriptive des variables clés :

- `catr` : la majorité des accidents surviennent sur des routes urbaines (catégories 3 et 4).
- `surf` :
  - 55 970 cas sur chaussée sèche (1),
  - 13 930 cas sur chaussée mouillée (2).
- `vma` : principales valeurs à 30, 50, 80 et 90 km/h, correspondant aux vitesses réglementaires usuelles.

### 3.2 Table Usagers

Cette table regroupe l'ensemble des attributs décrivant l'état des usagers pendant et après l'accident. De nombreuses variables peuvent avoir des valeurs codées comme "manquantes" ou "inconnues". De plus, certaines valeurs manquantes sont codées de plusieurs manières (par exemple, la variable `actp` peut prendre la valeur -1 pour "non renseigné", 0 ou B pour "inconnue").

#### Problèmes détectés et solutions apportées

- La variable `an_nais` comporte 2 598 valeurs NaN (2% des données). Nous avons transformé cette variable en âge au moment de l'accident.
- Les variables correspondant aux dispositifs de sécurité utilisés lors de l'accident contiennent entre 13% et 99% de valeurs NaN. Elles ne seront pas conservées pour le clustering.

### 3.3 Table Véhicules

La table Véhicules contient 11 attributs différents. Parmi eux, les variables `senc`, `obs`, `obsm`, `choc`, `manv` et `motor` peuvent avoir des valeurs manquantes codées par -1. De plus, `senc`, `catv`, `manv` et `motor` peuvent avoir des valeurs manquantes codées par 0.

#### Problèmes détectés

- Après vérification des colonnes de la table, tous les attributs comptent 93 585 valeurs, à l'exception de `occutc`, qui n'en a que 838. Une analyse rapide montre que cette colonne contient 99% de valeurs NaN. Nous avons donc décidé de la supprimer.

- La colonne `catv`, représentant la catégorie du véhicule impliqué dans l'accident, comporte 99 valeurs catégorielles différentes. Une dizaine de ces valeurs sont obsolètes depuis plusieurs années. Après analyse des proportions, nous avons regroupé les valeurs les plus proches afin de réduire le nombre de catégories. Nous avons appliqué la même méthode aux variables `choc` et `manv`.
- Les attributs concernant les collisions (objets mobiles ou fixes) sont très mal distribués (`obsm` : véhicule 70%, Aucun 19%, Piéton 9% ; `obs` : Aucun 84,9%). Nous avons donc décidé de binariser `obs` (`obstacle_fixe` : oui/non) et de remplacer `obsm` par deux colonnes binaires : `collision_vehicule_mb` et `collision_pieton`.

### 3.4 Table Caractéristiques

Cette table décrit les conditions générales des accidents de la route en France à travers des variables telles que :

- Num\_Acc
- atm (conditions atmosphériques)
- lum (conditions de luminosité)
- agg (type d'agglomération)
- lat, long (coordonnées géographiques), etc.

#### Vérification des valeurs manquantes

La seule colonne présentant des valeurs manquantes significatives est `adr` (adresse), avec **1 389 valeurs manquantes** ( 2,5 % des lignes). Ces valeurs manquantes ne sont pas problématiques pour les analyses statistiques, car les autres colonnes contiennent toutes les informations essentielles.

#### Analyse statistique et détection d'anomalies

L'analyse descriptive montre que :

- certaines variables codées contiennent des valeurs aberrantes égales à -1, par exemple dans `atm`, `lum` et `int` ;
- ces valeurs correspondent généralement à des données non renseignées ou inconnues et devront être remplacées par des valeurs manquantes (`NaN`) ou exclues selon le contexte.

### 3.5 Fusion des données

Les fichiers `caract-2023.csv` et `lieux-2023.csv` contiennent des informations complémentaires sur les mêmes accidents. Pour analyser les facteurs d'accidents dans leur globalité, il a été nécessaire de fusionner

ces deux jeux de données autour de leur clé commune Num\_Acc, identifiant unique de chaque accident. Nous avons fait de même avec les fichiers `vehicules-2023.csv` et `usagers-2023.csv`. La fusion a produit un grand nombre de colonnes. Pour faciliter les analyses, seules les variables jugées essentielles ont été conservées :

```
cols = ['atm', 'lum', 'agg', 'int', 'catr',
        'surf', 'vma', 'lat',
        'long', 'catv', 'manv', 'choc', 'obstacle_fixe', 'collision_véhicule_mb']
```

### Variables retenues et justification

Variable	Signification	Intérêt principal
atm	Conditions atmosphériques	Facteur environnemental majeur
lum	Niveau de luminosité	Influence directe sur la visibilité
agg	Type d'agglomération	Distingue zone urbaine / rurale
int	Type d'intersection	Nature de la configuration routière
catr	Catégorie de route	Niveau de dangerosité potentiel
surf	Etat de la surface	Lien direct avec l'adhérence
vma	Vitesse maximale autorisée	Mesure du contexte de circulation
lat, long	Coordonnées géographiques	Permettent la cartographie et l'analyse spatiale
catv	Type de véhicule	Distinguer les accidents impliquant des usagers vulnérables (motos, vélos) de ceux impliquant des véhicules lourds.
manv	Manoeuvre	Traduit directement le comportement de conduite
choc	Type de choc	Donne une information sur la dynamique de l'accident
obstacle_fixe	Choc avec obstacle fixe	Permet de distinguer les accidents de type « sortie de route » des collisions entre véhicules.
collision_véhicule_mb	Collision avec un véhicule mobile	Regrouper les accidents selon leur configuration principale
collision_pieton	Collision avec un piéton	Regrouper les accidents selon leur configuration principale
catu	Catégorie de l'usager	Identification des usagers les plus sensibles
grav	Gravité de l'accident	Quantifier la gravité de l'accident
sex	Sexe de l'usager	Analyse des comportements au volant
age	Age de l'usager	Analyse des comportements au volant
trajet	Type de trajet	Identifier le type de trajet

TABLE 2 – Variables retenues après fusion et leur justification.

### 3.6 Nettoyage des données

Après la fusion des fichiers , certaines variables contenaient des valeurs manquantes ou non renseignées. L'objectif de cette étape est de :

- Identifier ces anomalies,
- Les corriger ou les exclure,
- Garantir la validité des coordonnées géographiques pour les analyses spatiales futures.
- Analyser la corrélation entre variables
- Créer une variable gravite\_moyenne

#### Identification des valeurs manquantes ou codées

Lors de l'exploration des colonnes principales, plusieurs d'entre elles contenaient la valeur -1 (ou autre), utilisée comme code pour « non renseigné » :

Les colonnes concernées ont été nettoyées en remplaçant les valeurs -1 par des valeurs manquantes (NaN), afin qu'elles soient correctement interprétées lors des analyses statistiques.

#### Nettoyage des coordonnées géographiques

##### Problème rencontré

Les colonnes lat et long contenaient :

- des **virgules** comme séparateur décimal au lieu du point (48,8566 → 48.8566),
- des **valeurs incohérentes** situées en dehors du territoire français (ex. -17, -150).

##### Démarche de correction

Les traitements suivants ont été appliqués :

- Remplacer les virgules par des points,
- Conversion en type numérique (`float64`).

Cette opération assure que les coordonnées sont bien reconnues comme valeurs numériques exploitables.

**Vérification des valeurs extrêmes** Les valeurs minimales et maximales avant filtrage étaient :

- **Latitude** : -23.372504 → 51.04749
- **Longitude** : -176.207 → 168.09567

Ces bornes révèlent la présence de points situés en dehors de la France métropolitaine.

##### Filtrage géographique : France métropolitaine uniquement

Pour se concentrer sur le territoire national (hors DOM-TOM), les conditions suivantes ont été appliquées :

- Latitude conservée entre **41° et 52°**,
- Longitude conservée entre **-6° et 10°**.

Ces bornes couvrent l'ensemble de la France métropolitaine, des Pyrénées au Nord et de la Bretagne à l'Alsace.

**Résultat du filtrage :** Dimensions après filtrage géographique : (154 550, 21) Seules les observations correspondant à des positions valides en métropole ont été conservées, garantissant la **pertinence des analyses spatiales futures** (cartes de densité, clusters d'accidents, etc.).

### Analyse de corrélation

Nous avons analysé la corrélation entre les variables via une matrice de Cramer.

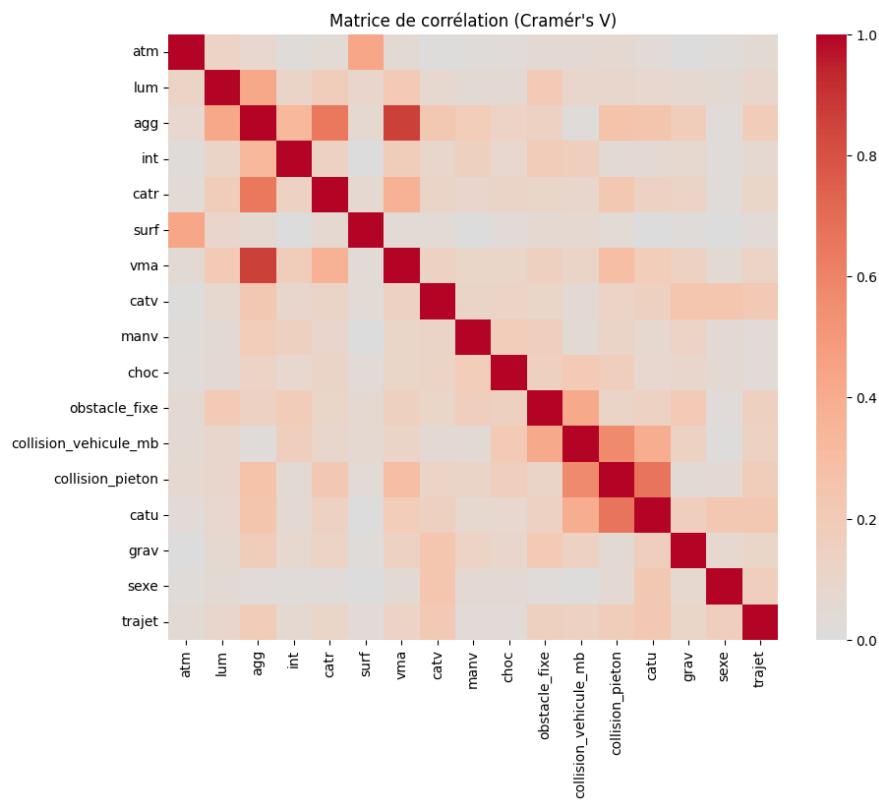


FIGURE 1 – Matrice de corrélation

Suite à ces résultats, nous avons supprimé :

- la variable **agg**, qui avait une corrélation supérieure à 70% avec **vma** et **catr**,
- la variable **catu**, également fortement corrélée avec **collision\_pieton**.

### 3.6.1 Creation variable gravite moyenne

## 3.7 Creation de la variable *gravite moyenne*

La variable *grav* issue du fichier des usagers n’tait pas directement exploitable dans le cadre du *clustering*. En effet, les modalites de gravite (indemne, blesse lger, blesse hospitalis, tu) etaient codes par des entiers ne respectant pas un ordre croissant de serite. Afin d’obtenir une representation plus coherente, nous avons proced ´a un **re-mapping ordinal** des valeurs. Les correspondances retenues sont les suivantes :

Valeur	Valeur initiale	Valeur remappee
Indemne	1	1
Tu	2	4
Blesse hospitalis	3	3
Blesse lger	4	2

Ce nouveau codage permet de refleter le niveau croissant de gravite des consequences d’un accident.

Nous avons ensuite agrege cette variable par numero d’accident (*Num\_Acc*) pour calculer une **gravite moyenne par accident**, notee *gravite\_moyenne*. Cette agregation permet de caracteriser chaque accident selon la serite moyenne des victimes impliques, tout en reduisant la redondance entre les enregistrements individuels d’usagers.

## 4 Clustering

### 4.1 Analyse et Interpretation du Clustering K-Means Adapte

#### 4.1.1 Choix de la methode

Nous avons utilise la methode **K-Means adaptee**. Le K-Means classique repose sur la minimisation de distances euclidiennes, ce qui le rend inadapte pour des donnees comportant des variables binaires ou catgorielles.

Pour corriger cela, nous avons mis en œuvre une version modifiee qui combine :

- une distance **euclidienne** pour les variables continues (*vma*, *lat*, *long*, *age*, *gravite\_moyenne*);
- une distance de **Hamming** pour les variables binaires ou catgorielles encodes (*conditions meteo*, *surface*, *manoeuvre*, *type de vhicule*, etc.).

Cette approche preserve la logique du K-Means tout en respectant la nature hetogene des donnees.

#### 4.1.2 Modélisation du K-Means adapté

L'algorithme implémenté suit les étapes suivantes :

1. Initialisation aléatoire de  $k$  centres mixtes (numériques et binaires)
2. Attribution des observations au centre le plus proche selon la distance mixte
3. Mise à jour des centres par la **moyenne** (pour les variables continues) et le **mode** (pour les binaires)
4. Répétition jusqu'à convergence

Nous avons testé plusieurs valeurs :

$$k \in \{3, 4, 5\}, \quad \alpha = 0.5$$

Résultats et évaluation : le meilleur modèle est obtenu avec :

$$k = 4 \quad \text{et} \quad \alpha = 0.5$$

Avec score silhouette : 0.095 → Bien que ce score reste faible, il demeure cohérent pour des données mixtes très hétérogènes combinant des variables continues, binaires et catégorielles. Il traduit une structure partielle mais réelle dans les données, c'est-à-dire que les clusters ne sont pas parfaitement séparés mais présentent des zones de cohérence interne. Cette segmentation est confirmée par la visualisation PCA, qui montre des regroupements distincts et interprétables malgré quelques chevauchements entre classes.

#### 4.1.3 Évaluation de la qualité des clusters et analyse des paramètres du K-Means adapté

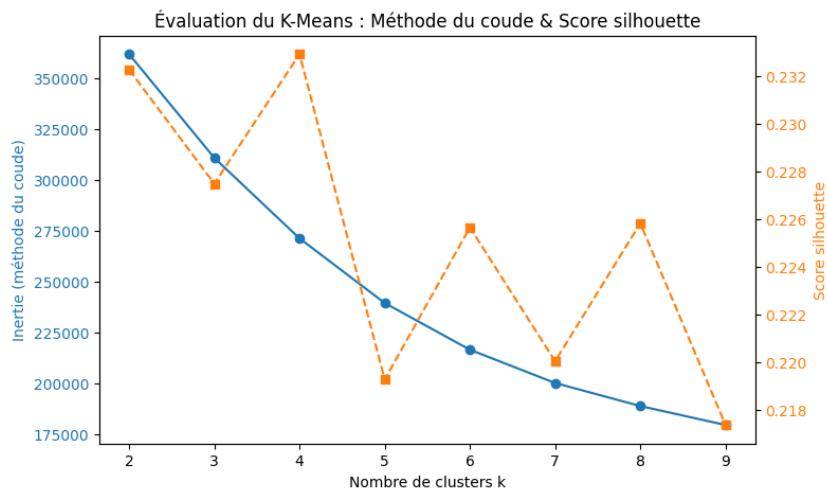


FIGURE 2 – Évaluation du K-Means : méthode du coude et score silhouette.

La qualité du regroupement a été évaluée à l'aide de la méthode du coude et du score silhouette. La courbe du coude met en évidence une forte diminution de l'inertie jusqu'à  $k = 4$ , suivie d'une stabilisation progressive : ce point représente donc le meilleur compromis entre la compacité interne des groupes et la complexité du modèle.

En parallèle, le score silhouette, dont les valeurs oscillent entre 0.21 et 0.23, atteint son maximum autour de  $k = 4$ . Cela confirme que cette configuration offre une séparation optimale entre les clusters, malgré la nature mixte des données. Ainsi, le modèle retenu repose sur quatre clusters ( $k = 4$ ,  $\gamma = 0.5$ ), ce qui correspond au meilleur équilibre entre cohérence interne et différenciation externe.

Le score silhouette global ( 0.095), bien que modéré, demeure satisfaisant compte tenu de la diversité et de l'hétérogénéité des variables (continues, binaires et catégorielles). Il traduit une structure partiellement marquée mais réelle, mise en évidence également par la projection PCA.

La normalisation des variables continues s'est révélée indispensable : sans cette étape, la vitesse et la position géographique auraient dominé le calcul des distances, faussant la répartition des observations. Malgré un certain chevauchement entre les groupes, les clusters obtenus restent stables, homogènes et interprétables, illustrant la capacité du K-Means adapté à identifier des profils distincts d'accidents routiers.

Ces résultats confirment que la segmentation en quatre clusters offre une représentation pertinente et exploitable des différents contextes d'accidents observés, tout en maintenant un bon compromis entre précision descriptive et robustesse statistique.

#### 4.1.4 Analyse des variables continues

Cluster	Vitesse (vma)	Latitude	Longitude	Âge	Gravité moyenne	Interprétation
0	0.57	-0.20	-0.04	0.11	2.29	Accidents modérément graves, majoritairement urbains
1	1.31	0.32	-0.37	-0.09	-0.26	Accidents à haute vitesse, peu graves
2	-0.71	0.61	-0.34	-0.03	-0.35	Accidents à basse vitesse en zone communale
3	-0.08	-1.13	0.83	0.07	-0.12	Accidents de jour dispersés, gravité faible

TABLE 3 – Analyse des variables continues par cluster.

#### 4.1.5 Analyse des variables catégorielles dominantes

Cluster	Principales modalités observées	Profil global
0	<code>surf_1.0</code> (route sèche), <code>atm_1.0</code> (temps clair), <code>choc_Avant</code> , <code>int_1.0</code> (hors intersection), <code>sexe_1.0</code> (homme), <code>trajet_5.0</code> (loisirs)	Accidents modérément graves en conditions normales
1	<code>surf_1.0</code> , <code>atm_1.0</code> , <code>int_1.0</code> (hors intersection), <code>catv_Voiture particulière</code> , <code>manv_Circulation normale</code> , <code>lum_1.0</code> (plein jour)	Accidents à vitesse élevée, hors agglomération
2	<code>catr_4</code> (voie communale), <code>atm_1.0</code> , <code>surf_1.0</code> , <code>choc_Avant</code> , <code>catv_Voiture particulière</code>	Accidents urbains à faible vitesse
3	<code>surf_1.0</code> , <code>atm_1.0</code> , <code>catv_Voiture particulière</code> , <code>lum_1.0</code> (plein jour), <code>trajet_5.0</code> (loisirs), <code>sexe_1.0</code> (homme)	Accidents routiers classiques en journée, conditions stables

TABLE 4 – Variables catégorielles dominantes par cluster.

#### 4.1.6 Interprétation générale

Cette segmentation met en évidence 4 profils d'accidents distincts :

- **Cluster 0** : accidents modérément graves, souvent frontaux, sur routes sèches et par temps clair, impliquant majoritairement des hommes au volant de voitures particulières.
- **Cluster 1** : accidents à haute vitesse, sur routes dégagées et hors intersection, généralement hors agglomération, associés à des trajets de loisirs.
- **Cluster 2** : accidents urbains à faible vitesse, majoritairement sur voies communales et en conditions normales (route sèche, bonne visibilité), avec une gravité faible.
- **Cluster 3** : accidents standards, survenus de jour et par temps clair, liés à des trajets personnels ou de loisirs, avec une gravité globalement faible à moyenne.

#### 4.1.7 Analyse détaillée des clusters

**Cluster 0** — Accidents modérément graves en conditions normales

- Conditions : principalement par temps clair (`atm_1.0`) et sur route sèche (`surf_1.0`), souvent hors intersection (`int_1.0`).
- Type de véhicule : majoritairement des voitures particulières (`catv_Voiture particulière`).
- Type de choc : essentiellement frontal (`choc_Avant`).
- Profil des conducteurs : majorité d'hommes (`sexe_1.0`) effectuant des trajets personnels ou de loisirs (`trajet_5.0`).
- Gravité : la plus élevée parmi les clusters ( $gravité_{moyenne} \approx 2.29$ ).
- Vitesse : modérée ( $vma \approx 0.57$ )
- Interprétation : ce cluster représente des accidents classiques de la circulation en conditions normales, souvent en zone urbaine ou

périurbaine, avec une gravité modérée à forte mais sans facteur météorologique aggravant.

**Cluster 1** — Accidents à haute vitesse, hors agglomération

- Conditions : routes dégagées et sèches (**surf\_1.0**), temps clair (**atm\_1.0**), hors intersection (**int\_1.0**).
- Contexte : liés à des trajets de loisirs (**trajet\_5.0**).
- Type de véhicule : principalement des voitures particulières (**catv\_Voiture particulière**).
- Manœuvre dominante : circulation normale (**manv\_Circulation normale**).
- Vitesse : la plus élevée parmi les clusters ( $vma \approx 1.31$ ).
- Gravité : relativement faible ( $gravitémoyenne \approx -0.26$ ).
- Interprétation : ces accidents surviennent surtout sur grands axes routiers à vitesse élevée, mais dans des conditions météorologiques favorables et sur des infrastructures sûres, expliquant la faible gravité observée.

**Cluster 2** — Accidents urbains à basse vitesse

- Conditions : sur voies communales (**catr\_4**), par temps clair (**atm\_1.0**) et sur route sèche (**surf\_1.0**).
- Type de choc : principalement frontal (**choc\_Avant**).
- Type de véhicule : voitures particulières (**catv\_Voiture particulière**).
- Vitesse : la plus faible parmi les clusters ( $vma \approx -0.71$ ).
- Gravité : basse ( $gravitémoyenne \approx -0.35$ ).
- Interprétation : ce cluster correspond à des collisions urbaines à faible vitesse, souvent lors de manœuvres ou d'interactions sur des voies à faible circulation. Ces accidents entraînent peu de blessés graves.

**Cluster 3** — Accidents de jour, trajets de loisirs

- Conditions : en plein jour (**lum\_1.0**), temps clair (**atm\_1.0**), routes sèches (**surf\_1.0**).
- Usagers : majoritairement des hommes (**sex\_1.0**) effectuant des trajets personnels ou de loisirs (**trajet\_5.0**).
- Type de véhicule : voitures particulières (**catv\_Voiture particulière**).
- Type de choc : principalement frontal (**choc\_Avant**).
- Vitesse : modérée ( $vma \approx -0.08$ ).
- Gravité : faible à moyenne ( $gravitémoyenne \approx -0.12$ ).
- Interprétation : ce groupe correspond à des accidents routiers standards, survenus en journée dans des conditions stables et prévisibles, souvent liés à des déplacements privés ou de loisirs.

#### 4.1.8 Visualisation (PCA)

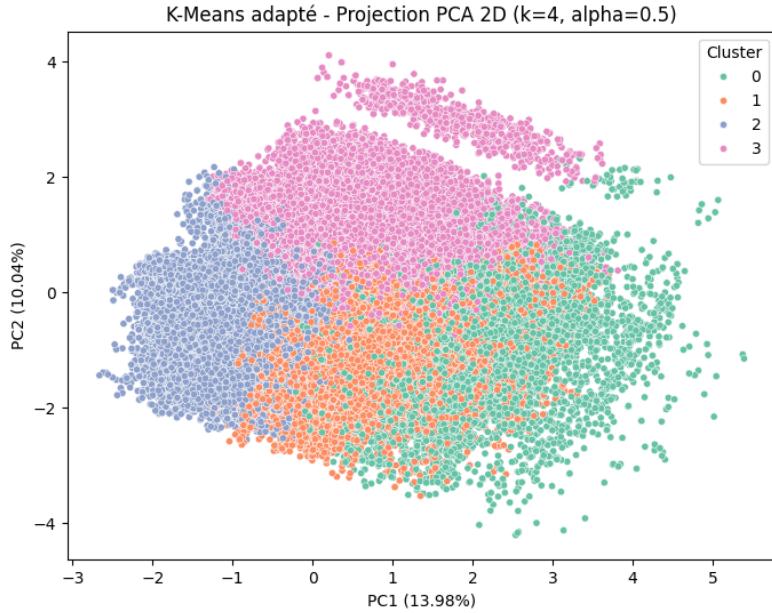


FIGURE 3 – K-Means adapté – Projection PCA 2D ( $k = 4$ ,  $\alpha = 0.5$ ).

La projection PCA 2D du K-Means adapté ( $k = 4$ ,  $\alpha = 0.5$ ) met en évidence **quatre ensembles principaux**. Les groupes apparaissent partiellement séparés, avec certaines zones de chevauchement, ce qui traduit la complexité et la continuité naturelle des contextes d'accidents. On observe néanmoins des régions de densité distinctes, correspondant à des profils spécifiques d'accidents :

- le **cluster rose** se distingue nettement dans la partie supérieure du graphique, suggérant un profil d'accidents bien caractérisé ;
- les **clusters vert et orange** présentent une légère superposition, traduisant une proximité dans leurs conditions d'occurrence ;
- le **cluster bleu** se positionne plus à gauche, formant une zone cohérente et compacte.

#### 4.1.9 Réduction de dimension et visualisation

Cette étape vise à améliorer la lisibilité des clusters en réduisant la dimensionnalité des données grâce à une analyse en composantes principales (PCA). La figure 4 montre quatre zones principales partiellement distinctes, confirmant la structure détectée par le K-Means adapté. Le chevauchement central traduit la proximité entre certains contextes d'accidents, tandis que les zones périphériques (rose et cyan) se démarquent davantage. Ainsi, la PCA permet de visualiser clairement la cohérence et la séparation globale des profils d'accidents identifiés.

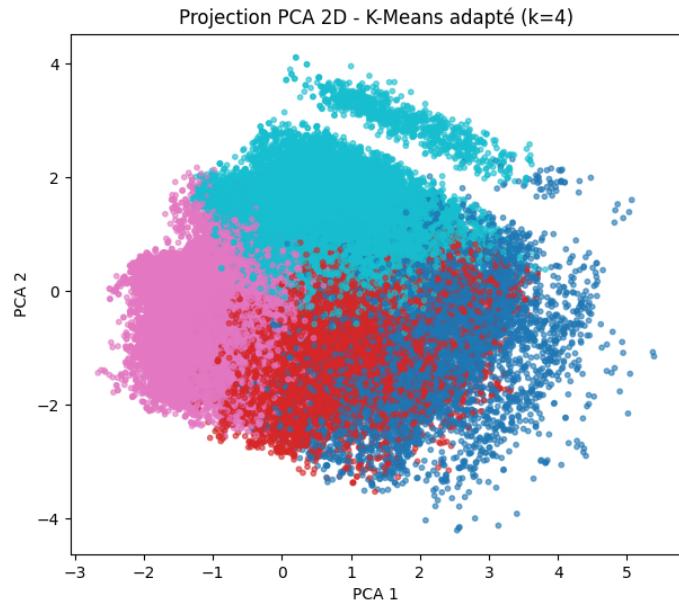


FIGURE 4 – Projection PCA 2D – K-Means adapté ( $k = 4$ ).

#### 4.1.10 Analyse de la distribution et de la répartition des clusters



FIGURE 5 – Distribution de la vitesse et de la gravité moyenne par cluster.

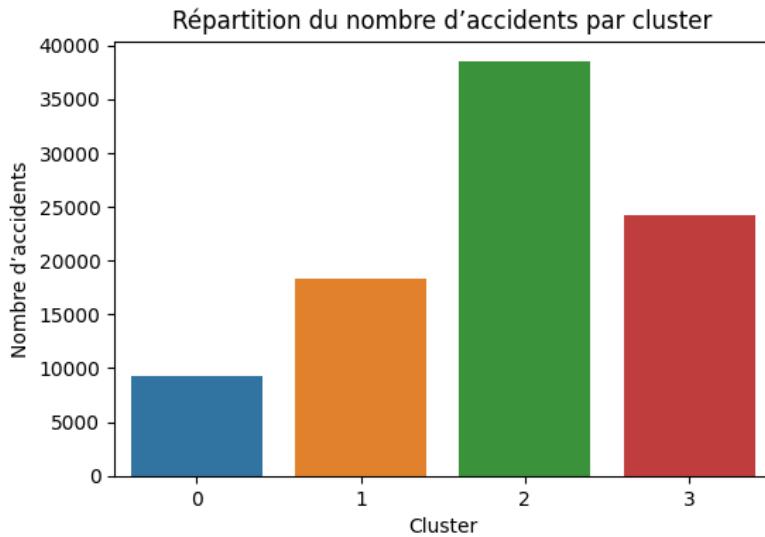


FIGURE 6 – Répartition du nombre d'accidents par cluster.

La distribution de la vitesse (vma) et de la gravité moyenne selon les clusters révèle des tendances marquées :

- Le cluster 0 présente les accidents les plus graves, associés à des vitesses modérées mais à des impacts plus sévères.
- Le cluster 1 regroupe des accidents à vitesse élevée, mais avec une gravité plus faible, probablement liés à des routes mieux aménagées.
- Les clusters 2 et 3 se distinguent par des vitesses faibles à moyennes et une gravité modérée, typiques des collisions urbaines ou des trajets de loisirs.

Le graphique de répartition montre que le cluster 2 concentre la majorité des accidents, confirmant la prédominance des situations urbaines à faible vitesse. Les clusters 0 et 1, moins fréquents mais plus graves, correspondent davantage à des accidents sur routes dégagées ou à chocs frontaux. Ces observations confirment la cohérence du K-Means adapté, où la vitesse, la gravité et le contexte routier structurent les différents profils d'accidents..

## 4.2 Méthodes spécialisées : K-Modes

Nous avons utilisé des méthodes de clustering spécialisées pour les données catégorielles ou mixtes. K-Modes est une extension de K-Means adaptée aux données purement catégorielles.

Au lieu d'utiliser la distance euclidienne, K-Modes s'appuie sur une mesure de dissimilarité simple : le nombre de catégories différentes entre deux individus (souvent appelée distance de Hamming). De plus, au lieu d'utiliser la moyenne pour représenter le centre d'un cluster, K-Modes utilise la mode (valeur la plus fréquente) pour chaque variable. Comme pour K-Means, il est nécessaire de choisir des centroïdes initiaux avant

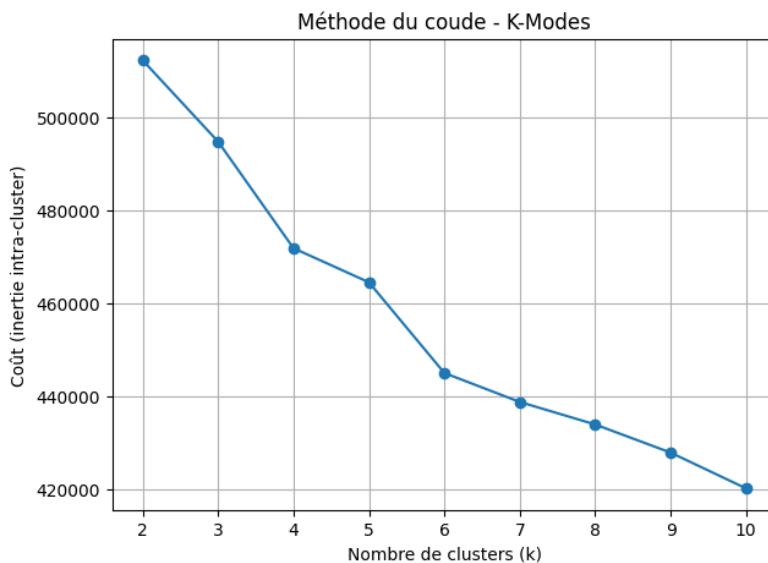
d’itérer. Deux méthodes principales d’initialisation existent : Huang (1997) et Cao (2009).

La méthode de Huang choisit les centroïdes initiaux aléatoirement parmi les points du jeu de données, en veillant à ce que les catégories les plus fréquentes soient bien représentées. Cette méthode est peu adaptée à notre dataset, car la présence de catégories très dominantes pourrait produire des clusters redondants. La méthode de Cao, quant à elle, repose sur la densité des points dans l’espace des attributs afin de choisir des centroïdes représentatifs, à la fois denses et éloignés les uns des autres. Cette approche est plus coûteuse en calcul mais plus robuste et fiable. La qualité du clustering dépend fortement de ce choix initial ; nous avons donc choisi la méthode de Cao, qui semble mieux adaptée à nos données. Nous avions envisagé de tester les deux méthodes et de comparer les scores de silhouette, mais le coût de calcul de cette dernière est trop élevé pour notre dataset.

Nos données ne sont pas exclusivement catégorielles : les attributs `age`, `vma`, etc. sont numériques. Nous avons donc créé une version des données entièrement catégorielle en transformant `age` et `vma` en tranches (0–18 ans, 65 ans et plus, etc.). Les autres attributs, plus difficiles à transformer ou moins pertinents, n’ont pas été conservés.

#### 4.2.1 Choix du nombre de clusters et de méthode

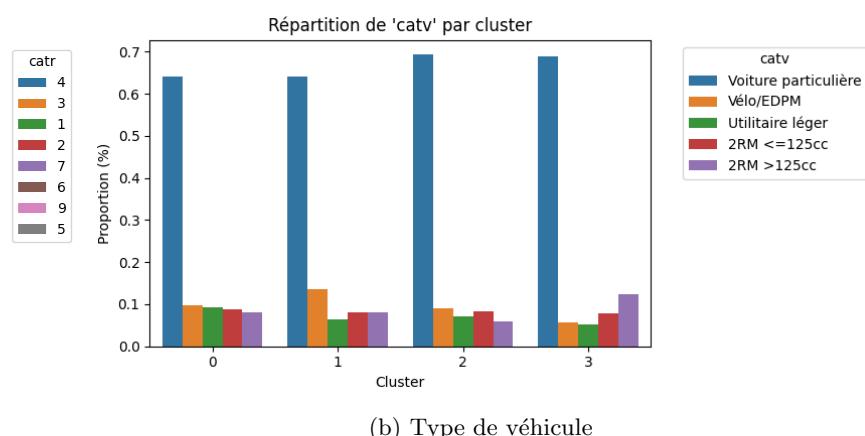
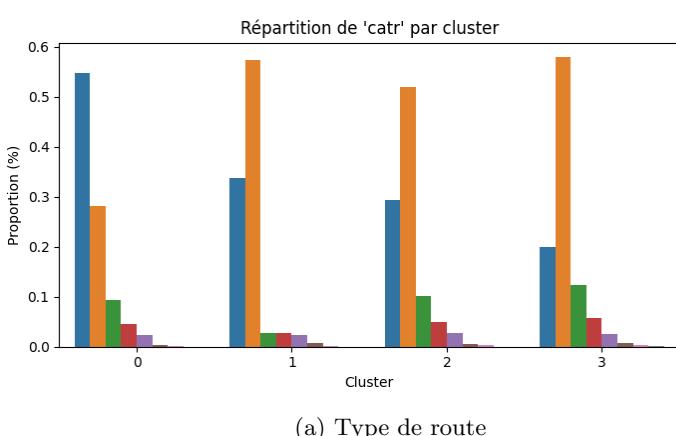
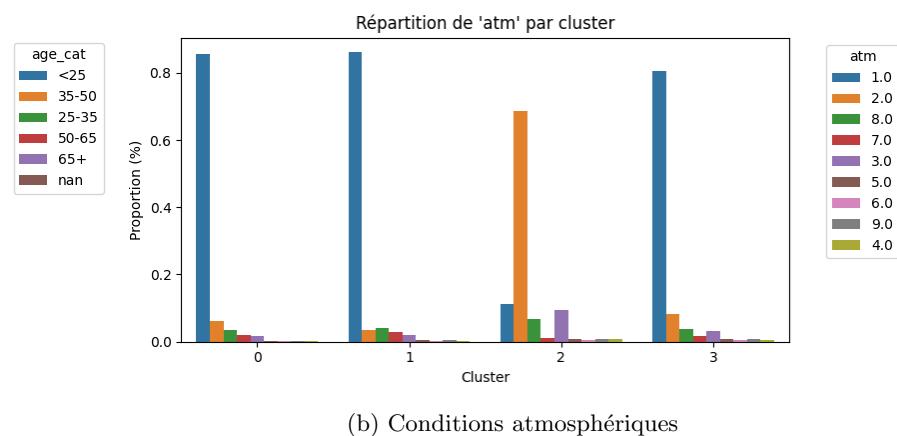
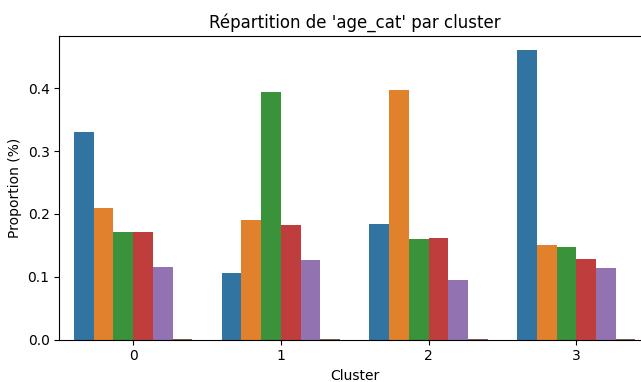
Pour déterminer le nombre de clusters, deux options s’offrent à nous : le score de silhouette ou le coût d’inertie (méthode du coude). En raison de la complexité du calcul du score de silhouette, celui-ci est trop long à exécuter. Nous nous sommes donc limités au coût d’inertie.

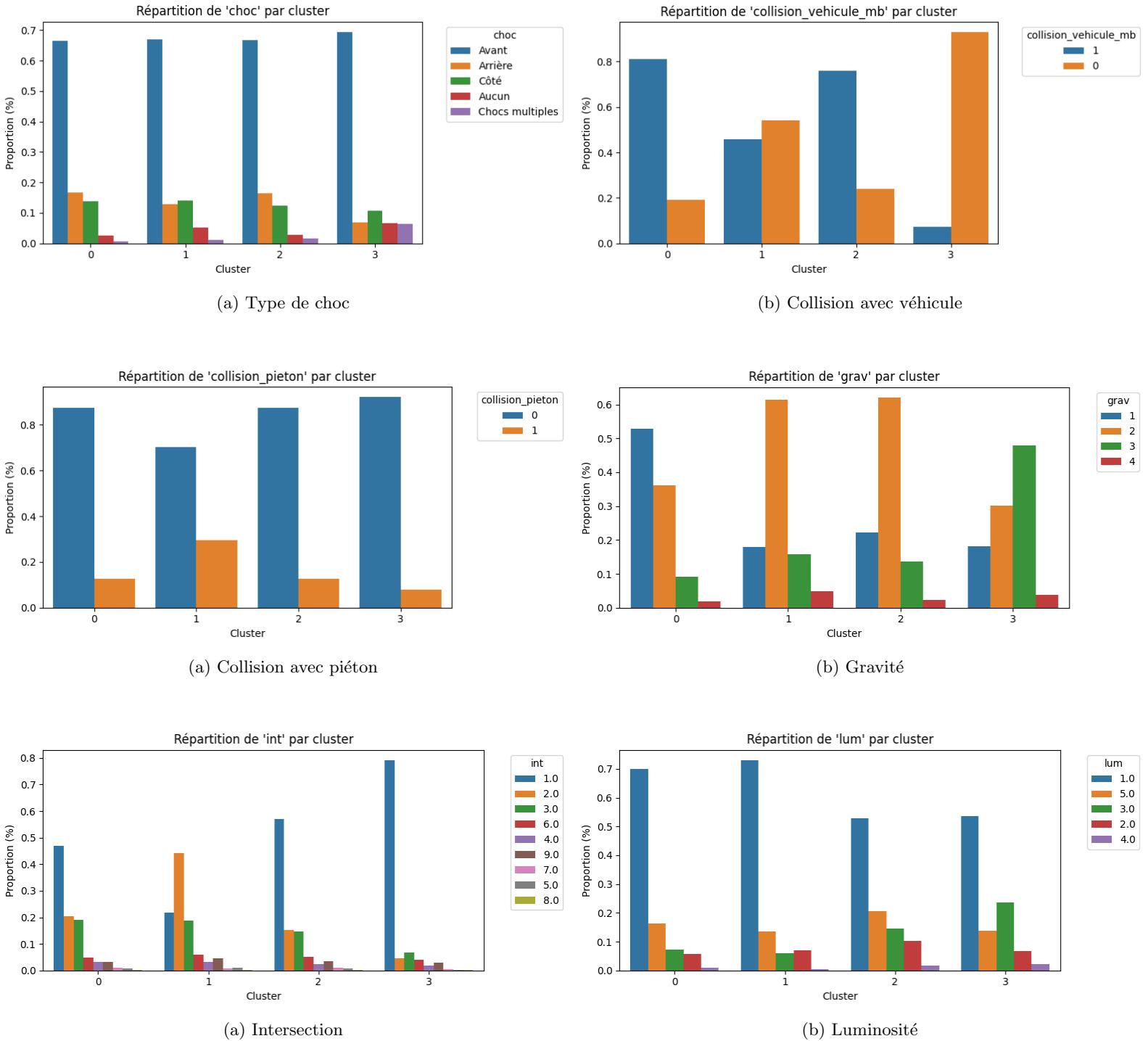


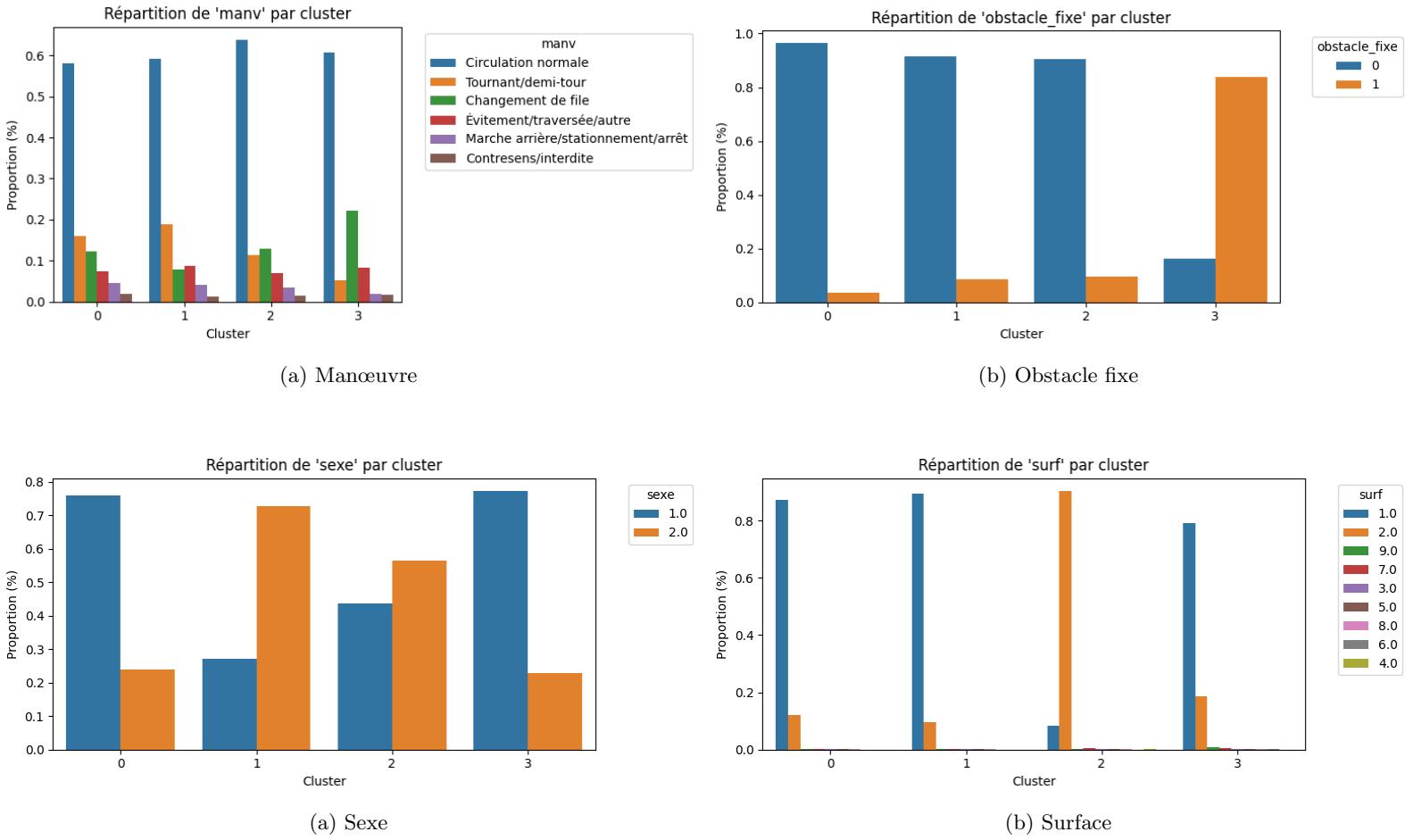
Visuellement, il est difficile d'identifier un "coude" sur la courbe, ce qui est un problème fréquent avec cette méthode. Nous avons donc calculé la pente de la courbe et choisi le  $k$  à partir duquel le taux de décroissance est inférieur à 5%. Cette méthode indique que le nombre optimal de clusters est de 4.

#### 4.2.2 Distribution des différentes catégories par cluster

La distribution des catégories par cluster permet de comprendre les caractéristiques de chaque groupe :







#### 4.2.3 Description des clusters

**Cluster 0 :** Accidents impliquant des usagers de tout âge, par temps ensoleillé en plein jour, sur chaussée sèche, sur des voies communales avec limitation de vitesse entre 25 et 50 km/h, hors intersection. Les collisions concernent principalement d'autres véhicules mobiles, avec des usagers majoritairement hommes, indemnes ou légèrement blessés, effectuant surtout des trajets de loisir. Ce cluster correspond aux accidents à faible gravité en agglomération à basse vitesse.

**Cluster 1 :** Accidents sur route départementale ou voie communale en intersection en X, par temps ensoleillé et chaussée sèche, avec limitation de vitesse 25–50 km/h. Environ la moitié des collisions impliquent un autre véhicule ou parfois un piéton. Les usagers ont 25 à 65 ans, majoritairement 25–35 ans, sont principalement des femmes effectuant des trajets domicile-travail et subissent des blessures légères.

**Cluster 2 :** Accidents par pluie légère en plein jour, sur chaussée

mouillée sur route départementale ou communale, limitation de vitesse 25–50 km/h, hors intersection, impliquant une collision avec un véhicule en mouvement. Les usagers sont équitablement hommes et femmes, âgés principalement de 35 à 50 ans, effectuant des trajets pour loisir ou domicile-travail et subissant de légères blessures.

**Cluster 3 :** Accidents sur chaussée sèche, par temps ensoleillé en plein jour ou la nuit, sans éclairage public, sur route départementale avec limitation de vitesse 25–80 km/h, hors intersection, impliquant une collision contre un obstacle fixe. Les usagers sont principalement des hommes de moins de 25 ans, effectuant un trajet de loisir, avec des blessures légères à graves nécessitant parfois une hospitalisation. Ce cluster correspond aux accidents graves contre un obstacle par de jeunes conducteurs.

### 4.3 Méthodes spécialisées : K-Prototypes

K-Prototypes est une extension naturelle de K-Means et K-Modes permettant de traiter simultanément des données mixtes, c'est-à-dire composées à la fois de variables numériques et catégorielles. L'algorithme combine les deux approches précédentes : la distance euclidienne est utilisée pour les attributs numériques, tandis qu'une mesure de dissimilarité basée sur la distance de Hamming est employée pour les variables catégorielles. Pour équilibrer l'influence respective de ces deux types de variables, un paramètre de pondération  $\gamma$  est introduit dans la fonction de coût.

Le principe reste similaire à celui de K-Means : on initialise un certain nombre de prototypes (centroïdes), puis on assigne chaque individu au cluster dont le prototype est le plus proche selon la distance mixte définie. Les prototypes sont ensuite mis à jour à chaque itération : la moyenne est recalculée pour les variables numériques et la mode est réévaluée pour les variables catégorielles. Le processus se répète jusqu'à convergence, c'est-à-dire lorsque les affectations de clusters ne changent plus de manière significative.

Cette méthode est particulièrement adaptée à notre cas, car notre jeu de données contient à la fois des variables continues (comme l'âge ou la vitesse maximale autorisée) et des variables catégorielles (comme le type de véhicule ou les conditions lumineuses). K-Prototypes permet donc une approche plus complète et réaliste du phénomène étudié, sans nécessiter de transformations excessives des données. Toutefois, comme pour K-Modes, le choix du nombre de clusters reste délicat.

#### 4.3.1 Choix du nombre de clusters

Pour les mêmes raisons que pour K-Modes, le nombre de clusters a été déterminé en fonction du coût d'inertie. Nous décidons de choisir 3

clusters.

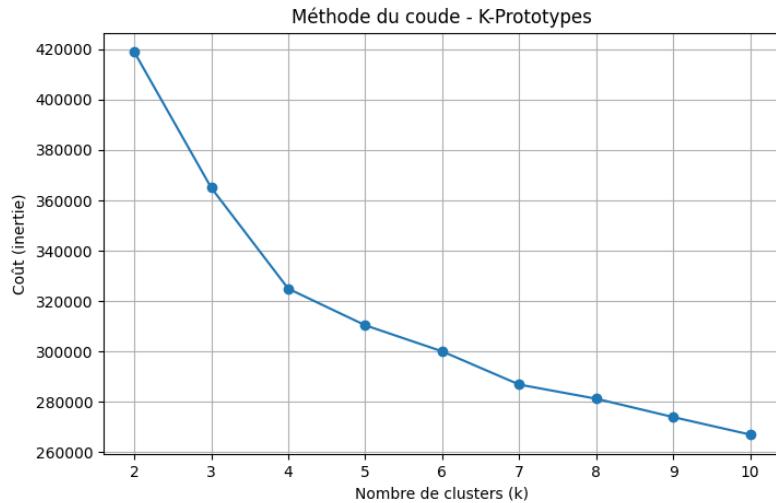
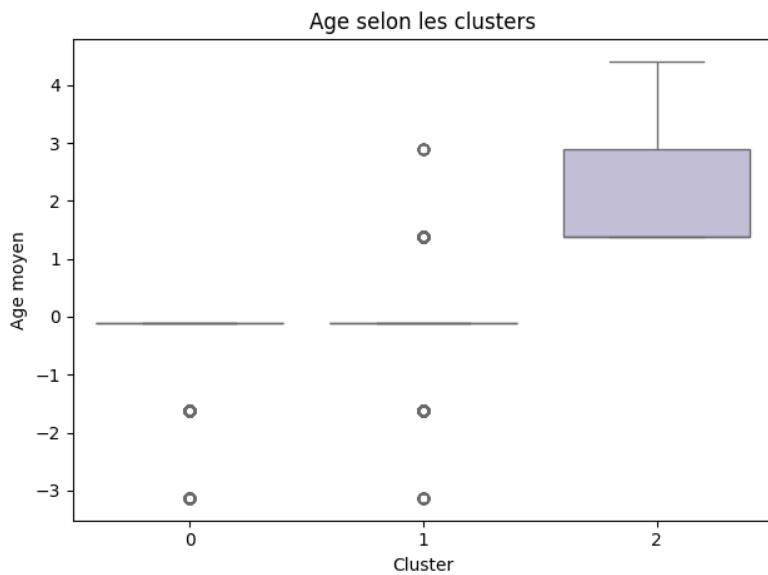
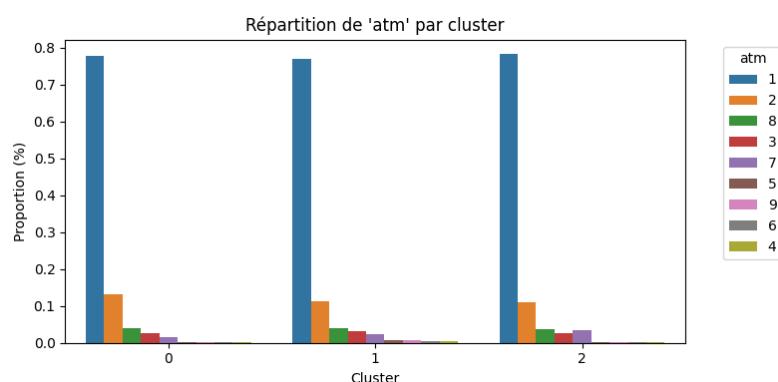


FIGURE 15 – Méthode du coude pour K-Prototypes

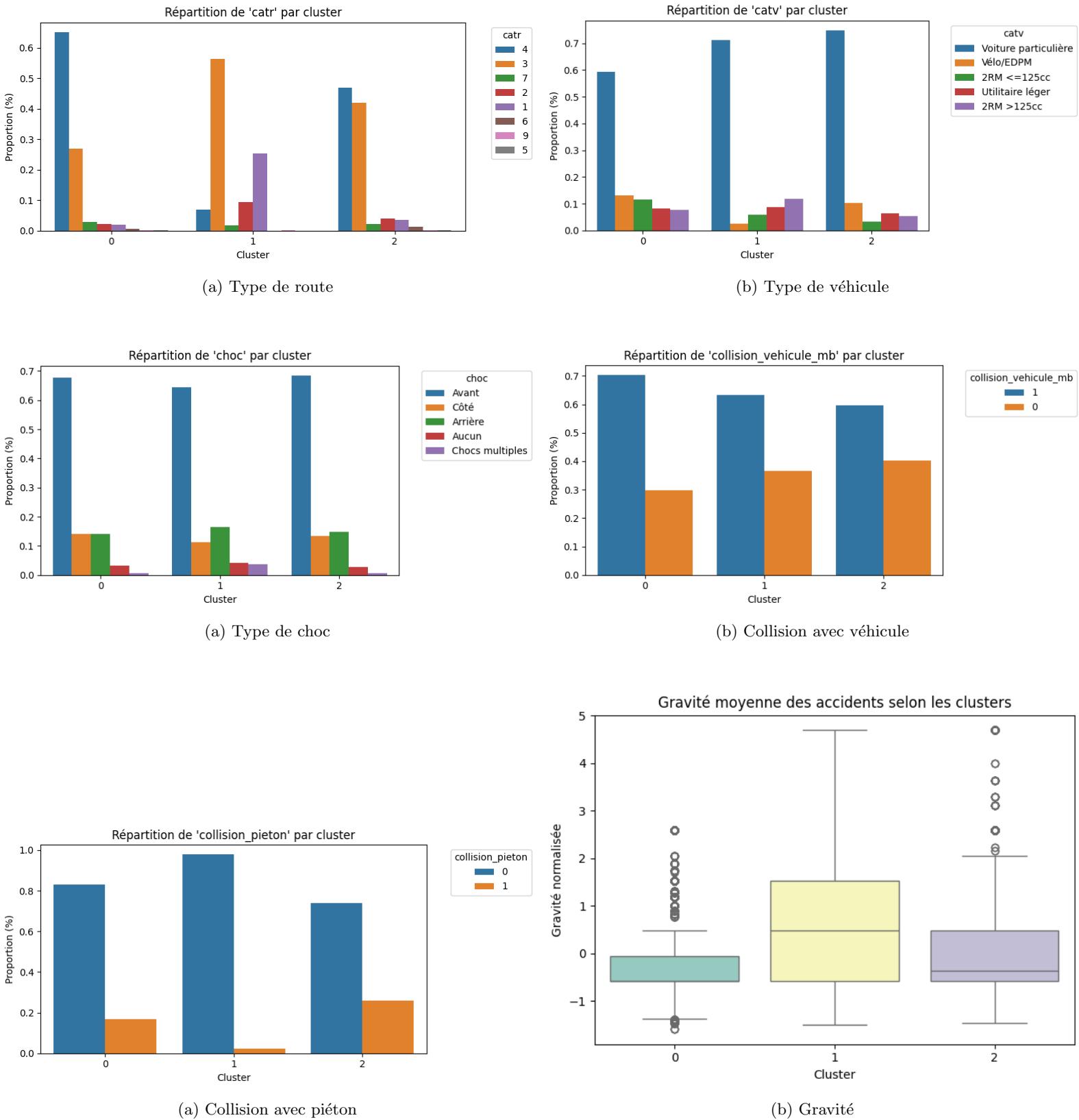
#### 4.3.2 Distribution des différentes catégories

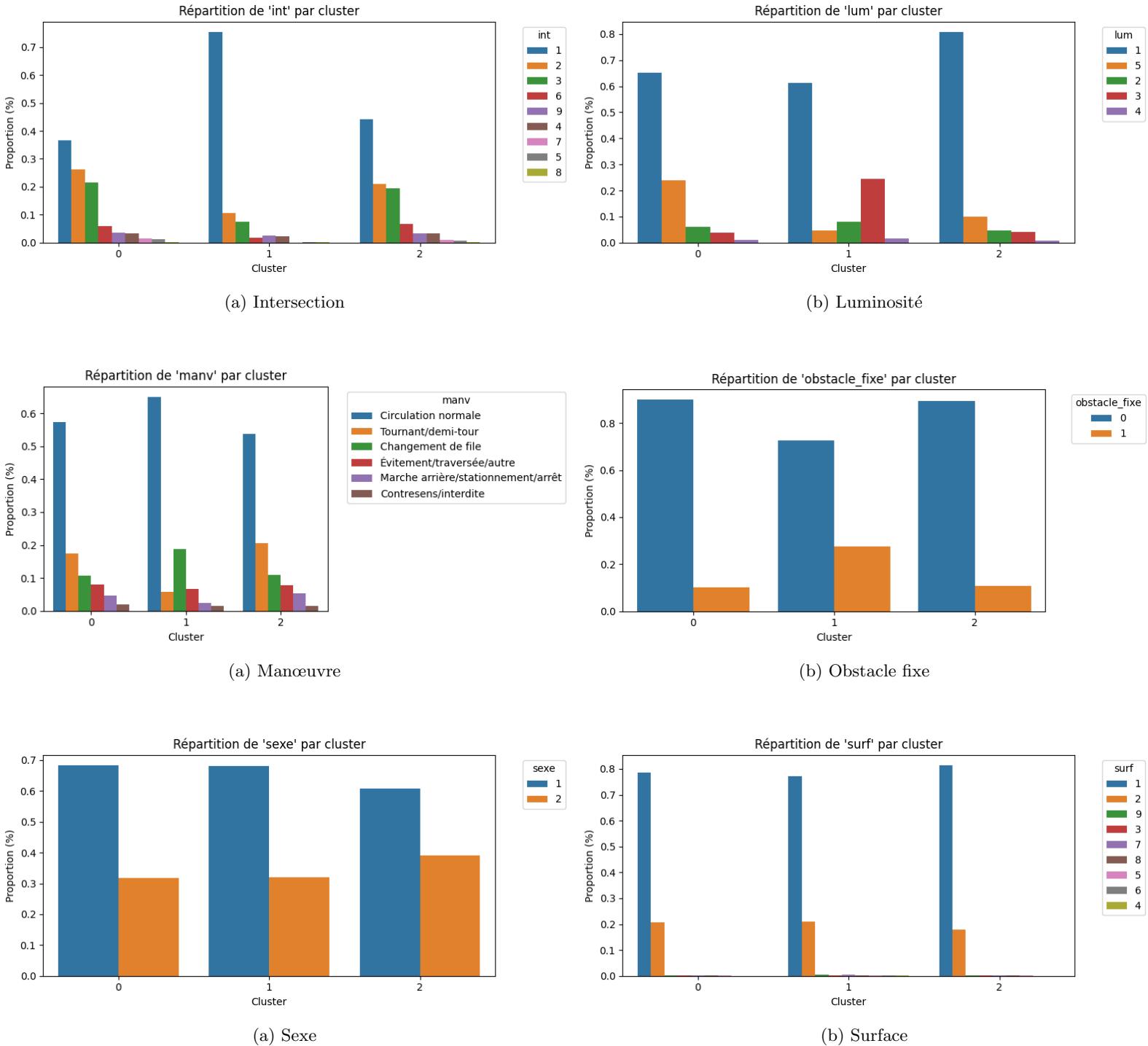


(a) Âge



(b) Conditions atmosphériques





### 4.3.3 Description des clusters

**Cluster 0 :** Accidents sur voies communales avec météo ensoleillée, principalement en plein jour, limitation de vitesse 50 km/h. Les collisions

concernent un autre véhicule hors intersection ou sur une intersection en X ou en T. Les usagers sont des hommes d'âge moyen ( 35 ans), et la gravité des accidents reste faible.

**Cluster 1 :** Accidents sur route départementale ou autoroute, chaussée sèche, météo ensoleillée, hors intersection, avec limitation de vitesse plus élevée. Les usagers sont majoritairement des hommes d'environ 35 ans effectuant un trajet de loisir. La gravité des accidents est plus importante que pour le cluster 0.

**Cluster 2 :** Accidents sur route départementale ou voie communale, chaussée sèche, météo ensoleillée, limitation de vitesse modérée. Les usagers sont hommes ou femmes, autour de 50 ans, effectuant des trajets de loisir. La gravité des accidents est faible.

## 5 Analyse des résultats

Les différents clustering réalisés montrent que la présence de variables fortement dominantes complique la distinction entre clusters. Par exemple, tous nos clusters représentent des accidents impliquant un véhicule particulier, en plein jour, sans manœuvre, sans collision avec un piéton et avec un choc avant, car un très grand pourcentage des données respectent ces caractéristiques.

Le fait d'avoir testé trois méthodes différentes nous permet de comparer leurs performances. Les méthodes K-Means, K-Modes et K-Prototypes partagent la même philosophie : partitionner les observations en  $k$  groupes selon un critère de minimisation de la distance intra-cluster. Cependant, leur comportement diffère selon la nature des variables, la métrique utilisée et la stratégie d'initialisation.

**K-Means** est adapté aux données numériques continues et bien séparables, mais devient inadapté dès que les variables sont qualitatives ou mixtes, même avec une adaptation par distance de Hamming, car cette dernière ne capture pas toujours la complexité des relations entre catégories.

**K-Modes** est plus efficace pour les données purement catégorielles. Il est sensible à l'initialisation et aux déséquilibres de classes, ce qui peut produire des clusters redondants ou dominés par les catégories majoritaires.

**K-Prototypes** traite efficacement les données mixtes en combinant distance euclidienne et distance de Hamming, avec un paramètre de pondération  $\gamma$ . Il est plus robuste et flexible que les deux autres méthodes, mais dépend du choix de  $\gamma$  et de l'initialisation.

En résumé, K-Means convient aux données numériques, K-Modes aux variables catégorielles, et K-Prototypes constitue le compromis le plus robuste pour les jeux de données mixtes.

Au-delà des aspects théoriques, K-Modes a permis d'extraire des profils d'accidents très clairs malgré les attributs redondants : on retrouve un profil type du jeune conducteur, un profil type d'accident sous pluie, et un profil type d'accident en agglomération. Pour K-Prototypes, les clusters obtenus sont moins distincts et ne permettent pas d'identifier de profils types clairement exploitables de plus la différence de complexité rend le temps de calcul pour K-Prototypes beaucoup trop élevé (2h pour la recherche du nombre de cluster).

Concernant le K-Means adapté, il s'est révélé particulièrement utile pour analyser les données mixtes. Il a permis d'identifier des profils d'accidents cohérents malgré la diversité des variables. Cette approche a nécessité un travail de préparation conséquent (fusion, nettoyage, encodage, normalisation), mais elle illustre bien la capacité du clustering à faire émerger des tendances structurelles dans un ensemble de données complexes. Les clusters obtenus traduisent des différences significatives de gravité, de vitesse et de conditions de circulation, confirmant la pertinence de cette adaptation du K-Means aux données mixtes.

## 6 Conclusion

Ce projet nous a permis de travailler sur de vraies données et nous a fait découvrir les problématiques liées aux données mal distribuées, manquantes ou obsolètes. Nous avons également exploré le clustering sur des données catégorielles, ce qui nous a conduit à étudier l'adaptation de K-Means et l'utilisation de méthodes spécialisées dérivées.

De plus, ce type de projet pourrait contribuer à l'amélioration de la sécurité routière, car il permet d'identifier des profils types d'accidents et donc de cibler des mesures préventives.

Avec plus de temps, nous aurions pu réaliser un clustering par thématiques d'accidents (environnement, usagers, etc.) puis effectuer un clustering croisé. Nous aurions également souhaité explorer d'autres méthodes de calcul de distance, comme la métrique de Gower, pour améliorer la comparaison entre observations mixtes.