

|  |
| --- |
| Accidents routiers en France |
|  |
| 9 juin  Emmanuel GAUTIER  Rachelle LE HOUEDEC  Rémi THINEY |

**Sommaire**

[2 Présentation du projet 3](#_Toc168860925)

[2.1 Origine des données 3](#_Toc168860926)

[2.2 Description générale 3](#_Toc168860927)

[2.3 Méthodologie 4](#_Toc168860928)

[2.4 Avertissement : 4](#_Toc168860929)

[3 Description des données utilisées 6](#_Toc168860930)

[3.1 Regroupement des fichiers par rubriques 6](#_Toc168860931)

[3.2 Description du contenu des rubriques 6](#_Toc168860932)

[3.3 caracteristiques 6](#_Toc168860933)

[3.4 véhicules 6](#_Toc168860934)

[3.5 Lieux 6](#_Toc168860935)

[3.6 Rubrique « Usagers » 7](#_Toc168860936)

[3.6.1 Les variables des fichiers 7](#_Toc168860937)

[3.6.2 Exploration des données 10](#_Toc168860938)

# Présentation du projet

L’objectif principal de ce projet est d’essayer de prédire la gravité des accidents routiers en France. Les prédictions seront basées sur les données historiques.

2ème objectif à déterminer

## Origine des données

Les données utilisées dans cette étude proviennent de la base de données annuelles des accidents corporelles de la circulation routière en France de 2005 à 2019

<https://www.data.gouv.fr/en/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2019/#community-discussions>

## Description générale

Les fichiers de données analysés contiennent des informations sur les accidents corporels en France (y compris les DOM-TOM) entre 2005 et 2022, soit pendant 18 années.

Les informations des fichiers recensent les accidents corporels survenu sur une voie ouverte à la circulation publique, impliquant au moins un véhicule et ayant fait au moins une victime ayant nécessité des soins. Ces informations sont saisies par l’unité des forces de l’ordre (police, gendarmerie, etc.) qui est intervenue sur le lieu de l’accident.

Ces fichiers contiennent dont des informations de localisation de l’accident, des informations concernant les caractéristiques de l’accident, les véhicules impliqués et leurs victimes.

Les informations sont réparties dans 4 rubriques donnant lieu à 4 types de fichiers :

1. La rubrique CARACTERISTIQUES qui décrit les circonstances générales de l’accident

2. La rubrique LIEUX qui décrit le lieu principal de l’accident même si celui-ci s’est déroulé à une intersection

3. La rubrique VEHICULES impliqués

4. La rubrique USAGERS impliqués

il y a 1 fichiers au format CSV pour chaque rubrique et pour chaque année, soit un total de 72 fichiers CSV

Chacune des variables contenues dans une rubrique doit pouvoir être reliée aux variables des autres rubriques. Le n° **d'identifiant de l’accident** (Cf. "Num\_Acc") présent dans ces 4 rubriques permet d'établir un lien entre toutes les variables qui décrivent un accident.

Quand un accident comporte plusieurs véhicules, il faut aussi pouvoir relier chaque véhicule à ses occupants. Ce lien est fait par la variable **id\_vehicule**.

## Méthodologie

La 1ère étape du projet consistera à analyser les données, le contenu des différentes variables de la base de donnée (ex : distribution des données, identification des valeurs nulles, des données aberrantes etc.)

Une fois cette analyse effectué, la 2ème étape du projet consistera à déterminer les variables pertinentes, à nettoyer les données (traitement des valeurs nulles, décision sur des valeurs erronées (ex. certaines données avec des valeurs -1 et null), traitement des valeurs abérantes etc … afin d’avoir un jeu de donnée prêt à la modélisation

Enfin, la dernière étape consistera à déterminer le meilleur modèle permettant de répondre aux problématiques identifiées, à savoir

* Prédire la gravité des accidents
* ????

## Avertissement :

Au fil des années, les fichiers, les process de saisies ont évoluées, certaines données ne peuvent donc pas être comparés dans le temps :

* Les données sur la qualification de blessé hospitalisé depuis l’année 2018 ne peuvent être comparées aux années précédentes suite à des modifications de process de saisie des forces de l’ordre. L’indicateur « blessé hospitalisé » n’est plus labellisé par l’autorité de la statistique publique depuis 2019.
* A partir des données de 2021, les usagers en fuite ont été rajoutés, cela entraîne des manques d’informations sur ces derniers, notamment le sexe, l’âge, voire la gravité des blessures (indemne, blessé léger ou blessé hospitalisé).

# Description des données utilisées

## Regroupement des fichiers par rubriques

La répartition des données dans 72 fichiers nous impose de regrouper ces données.

Nous concaténons les données dans 4 rubriques, chaque rubrique contient alors les données de 18 ans.

Le changement de format en 2019 nécessite des conversions. TODO préciser.

L’ensemble 18 années des 4 rubriques précités seront par la suite regroupés dans un fichier unique. (à réaliser)

## Description du contenu des rubriques

Les nombres de colonnes sont ceux de 2005 à 2018 et de 2019 à 2022.

Les nombres d’observations sont ceux de 2005 à 2022.

Des répartitions par modalités sont.

De nombreuses valeurs catégorielles sont codées avec des entiers

Pour réaliser les statistiques exploratoires nous utilisons un outil qui automatise les statistiques exploratoires.

**caracteristiques**

Nombre d’observations : 1 176 873

**véhicules**

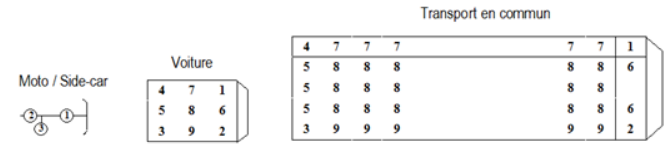
**Lieux**

**Rubrique « Usagers »**

* Les variables des fichiers

Le fichier « Usagers » contient les variables suivantes :

* Num\_Acc : Identifiant de l’accident identique à celui du fichier "rubrique CARACTERISTIQUES" repris pour chacun des usagers décrits impliqués dans l’accident
* id\_usager : Identifiant unique de l’usager (y compris les piétons qui sont rattachés aux véhicules qui les ont heurtés). Cette variable a été ajoutée aux fichiers en 2021.
* id\_vehicule : Identifiant unique du véhicule repris pour chacun des usagers occupant ce véhicule (y compris les piétons qui sont rattachés aux véhicules qui les ont heurtés). Cette variable a été ajoutée aux fichiers en 2019.
* num\_Veh : Identifiant du véhicule repris pour chacun des usagers occupant ce véhicule (y compris les piétons qui sont rattachés aux véhicules qui les ont heurtés)
* place : Permet de situer la place occupée dans le véhicule par l'usager au moment de l'accident. Les modalités vont de 1 à 9, l’illustration ci-dessous montre les différentes places possibles :



* catu : Catégorie d'usager. Les différentes modalités de cette variables sont les suivantes :
  + 1 – Conducteur
  + 2 – Passager
  + 3 – Piéton
  + 4 - Piéton en roller ou en trottinette (catégorie déplacée, à partir de l’année 2018, vers le fichier "Véhicules" Catégorie du véhicule : 99 - Autre véhicule. Cette catégorie est désormais considérée comme un véhicule : engin de déplacement personnel)
* grav : Gravité de blessure de l'usager, les usagers accidentés sont classés en trois catégories de victimes plus les indemnes :
  + 1 – Indemne
  + 2 – Tué
  + 3 – Blessé hospitalisé
  + 4 – Blessé léger
* sexe : Sexe de l'usager
  + 1 – Masculin
  + 2 – Féminin
* an\_nais : Année de naissance de l'usager
* trajet : Motif du déplacement au moment de l’accident :
  + -1 – Non renseigné
  + 0 – Non renseigné
  + 1 – Domicile – travail
  + 2 – Domicile – école
  + 3 – Courses – achats
  + 4 – Utilisation professionnelle
  + 5 – Promenade – loisirs
  + 9 – Autre
* sécu : Présence et utilisation des éléments de sécurités, codées sur 2 caractères. 1er caractère : existence d’un Équipement de sécurité
  + 1 – Ceinture
  + 2 – Casque
  + 3 – Dispositif enfants
  + 4 – Equipement réfléchissant
  + 9 – Autre

2ème caractère concerne l’utilisation de l’Équipement de sécurité

* + 1 – Oui
  + 2 – Non
  + 3 – Non déterminable

Cette variable a disparu en 2019

* secu1 : Variable présente à partir de 2019. Il s’agit de l’utilisation avec jusqu’à 3 équipements possibles pour un même usager (notamment pour les motocyclistes dont le port du casque et des gants est obligatoire). Le renseignement du caractère indique la présence et l’utilisation de l’équipement de sécurité
  + -1 – Non renseigné
  + 0 – Aucun équipement
  + 1 – Ceinture
  + 2 – Casque
  + 3 – Dispositif enfants
  + 4 – Gilet réfléchissant
  + 5 – Airbag (2RM/3RM)
  + 6 – Gants (2RM/3RM)
  + 7 – Gants + Airbag (2RM/3RM)
  + 8 – Non déterminable
  + 9 – Autre
* secu2 : idem SECU 1
* secu3 : idem SECU 1
* locp : Localisation du piéton
  + -1 – Non renseigné
  + 0 – Sans objet

Sur chaussée :

* + 1 – A + 50 m du passage piéton
  + 2 – A – 50 m du passage piéton

Sur passage piéton :

* + 3 – Sans signalisation lumineuse
  + 4 – Avec signalisation lumineuse

13 Divers :

* + 5 – Sur trottoir
  + 6 – Sur accotement
  + 7 – Sur refuge ou BAU
  + 8 – Sur contre allée
  + 9 – Inconnue
* actp : Action du piéton :
  + -1 – Non renseigné

Se déplaçant

* + 0 – Non renseigné ou sans objet
  + 1 – Sens véhicule heurtant
  + 2 – Sens inverse du véhicule

Divers

* + 3 – Traversant
  + 4 – Masqué
  + 5 – Jouant – courant
  + 6 – Avec animal
  + 9 – Autre
  + A – Monte/descend du véhicule
  + B – Inconnue
* etatp : Cette variable permet de préciser si le piéton accidenté était seul ou non :
  + -1 – Non renseigné
  + 1 – Seul
  + 2 – Accompagné
  + 3 – En groupe
* Exploration des données

La partie exploration des données a été effectué sur le fichier « usagers » regroupant les 18 fichiers CSV des usagers

Ce fichier contient donc l’ensemble des usagers impliqués dans un accident de 2005 à 2022

Le fichier total contient 2 636 377 lignes

La répartition par année la suivante :

|  |  |
| --- | --- |
| Année | Nb de lignes |
| 2005 | 197498 |
| 2006 | 187085 |
| 2007 | 188457 |
| 2008 | 170960 |
| 2009 | 165962 |
| 2010 | 154192 |
| 2011 | 148543 |
| 2012 | 138628 |
| 2013 | 128694 |
| 2014 | 132186 |
| 2015 | 130378 |
| 2016 | 133422 |
| 2017 | 136021 |
| 2018 | 130169 |
| 2019 | 132977 |
| 2020 | 105295 |
| 2021 | 129248 |
| 2022 | 126662 |
| Total | 2636377 |

On peut rapidement constater que le nombre d’usagers diminue fortement et de manière régulière de2005 à 2013. Le nombre d’usagers reste à peu près constant depuis 2013 à l’exception de l’année 2020

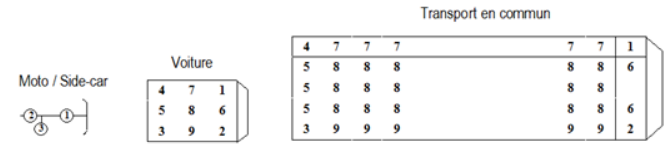
Pour l’année 2020, une explication du « trou » peut être la pandémie, qui a limité les déplacements durant une partie de l’année

Nous allons faire un zoom sur les variables explicatives du fichier « Usagers », à savoir les variables à l’exception des variables de type identifiants. On va donc explorer les variables

place, catu, grav, sexe, an\_nais, trajet, secu, secu1, secu2, secu3, locp, actp, etatp

* Variable place

La variable place est présente dans les fichiers de données depuis 2005. Il s’agit d’une variable catégorielle. Les modalités attendus sont de 1 à 9 (place dans le véhicule) selon le schéma



⇨ Estimation de la qualité du remplissage de cette données :

\*Détection du taux de valeurs nulles de la donnée :



On observe qu’il y a entre 8 à 9% de valeurs non renseignées entre 2009 et 2018. Il n’y a pas de valeurs non renseigné entre 2005 et 2008 nul après 2018. Est-ce ce cependant les données renseignés sont correctes ?

La modalité la plus représenté les la modalité « 1 »

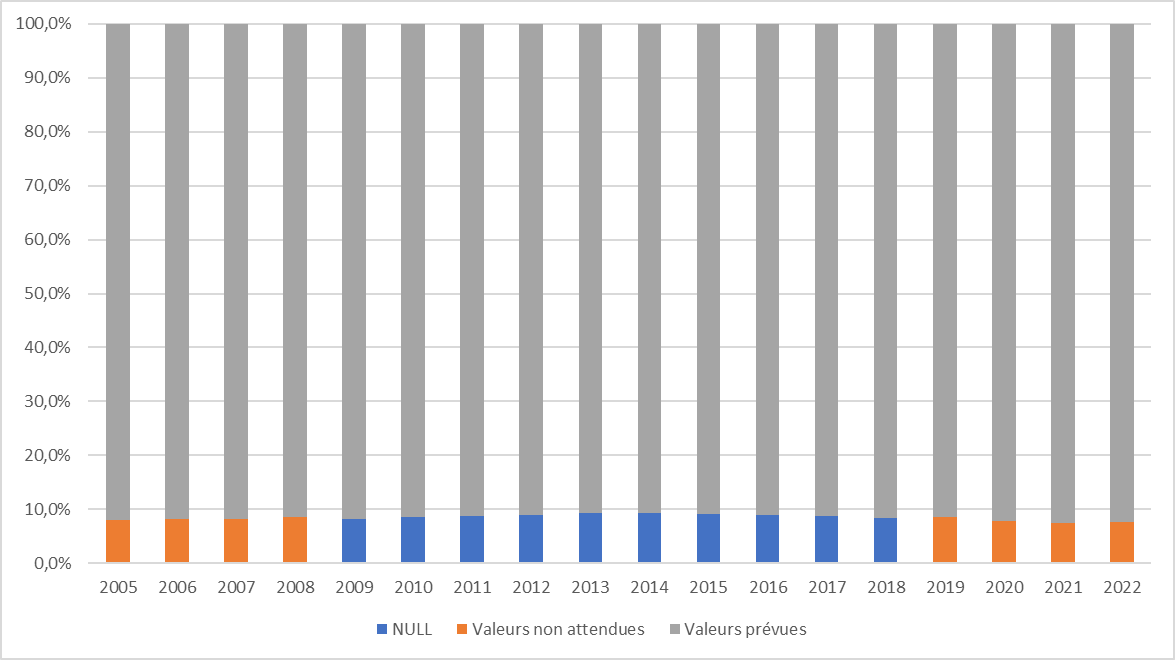
\* Répartition des modalités par année :





* + - En approfondissant l’analyse et en regardant la répartition des modalités dans le temps, on voit que on a des valeurs non conformes : « 0 », « -1 », « 10 » : selon des années, on a soit des « non renseignés » soit des « valeurs non conformes ». Cela correspond peut-être à des changements dans les consignes ou outils de saisies

Globalement, on a plus de 90 % des données conformes aux modalités attendues. Ce taux est constant dans le temps

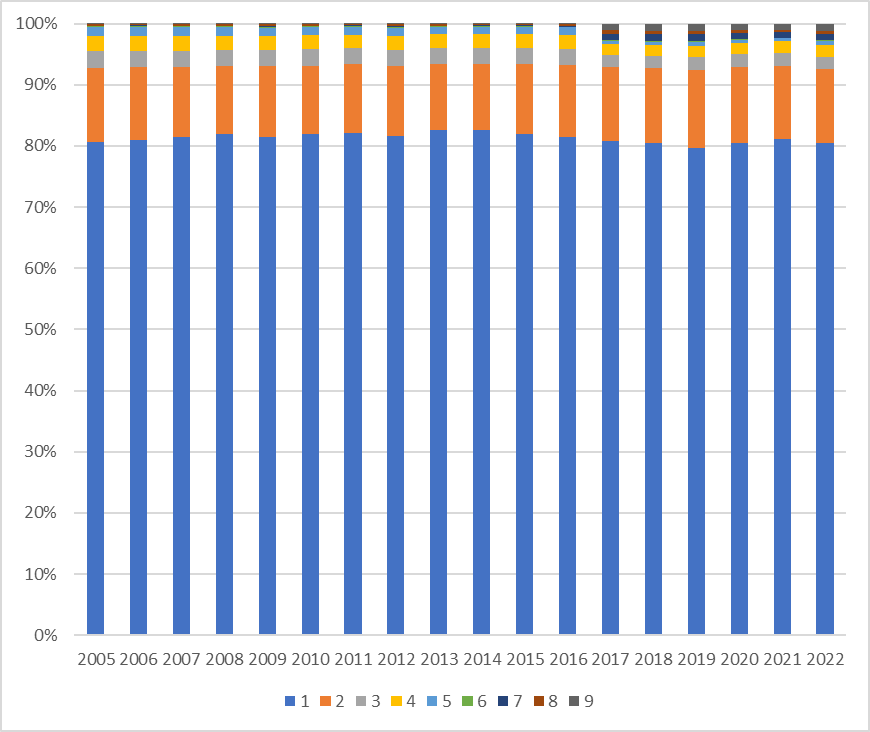


\*Synthèse :

Cette variable semble correctement renseigné dans plus de 90% des cas.

La question à se poser est de de savoir comment on traite les valeurs « non renseigné », « 0 », « -1 » et « 10 » qui représentent 8 à 10% des modalités de cette variable

⇨ 1ère analyse :



* + - On remarque que la modalité « 1 » qui représente la place du conducteur représente 80% des usagers
    - Les modalités « 1 » et « 2 » (emplacement passager à côté du conducteur) représentent plus de 90% des usagers
    - La proportion des différentes modalités varient peu dans le temps
* **Variable catu**

La variable catu représente la catégorie d'usager. Il s’agit d’une variable catégorielle. Les différentes modalités de cette variables sont les suivantes :

* + 1 – Conducteur
  + 2 – Passager
  + 3 – Piéton
  + 4 - Piéton en roller ou en trottinette (avant 2018)

⇨ Estimation de la qualité du remplissage de cette données :

\*Détection du taux de valeurs nulles de la donnée :



On observe qu’il n’y a jamais de valeur nulles

La modalité la plus représenté les la modalité « 1 »

\*Répartition des modalités par année :

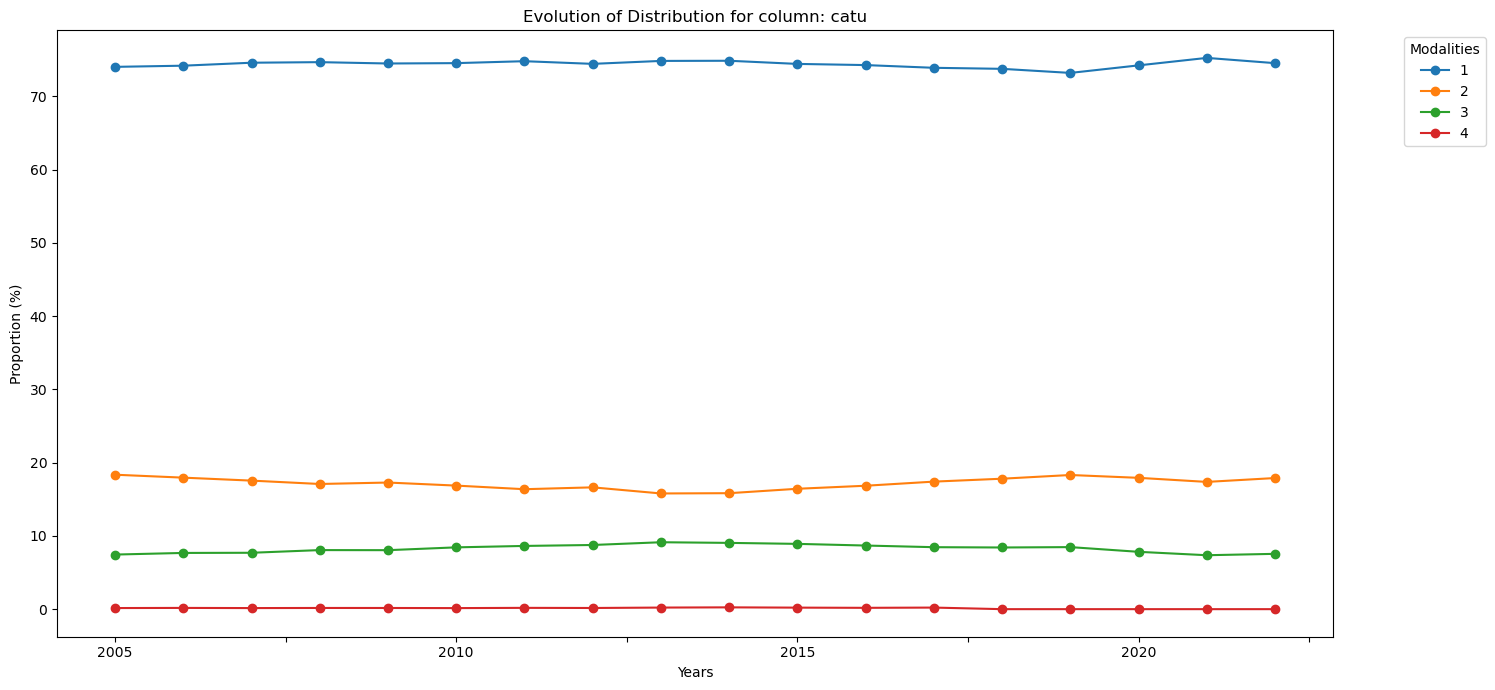
Les modalités sont conformes

La modalité « 4 » est présente de 2005 à 2017 conformément à ce qui a été décrit

\*Synthèse :

Cette variable est correctement renseigné et ne nécessite aucune action de nettoyage et de retraitement.

⇨ 1ère analyse



* + - On remarque que la modalité « 1 » qui représente le conducteur plus de 70% des usagers
    - Les modalités « 2 » (passager) passager à côté du conducteur) représentent plus de 90% des usagers
    - La proportion des différentes modalités varient peu dans le temps

Cette variable est parfaitement renseignée (pas de valeur non renseigné, pas de valeur aberrante, toutes les modalités correspondent à la liste de valeur indiqué)

On voit que la modalité « 1 » qui correspond au conducteur représente environ 75% des enregistrements, ce qui est cohérent avec la variable « place »

La modalité 4 est très faible voire négligeable (<1%) avant la suppression en 2018

Point à éclairer :

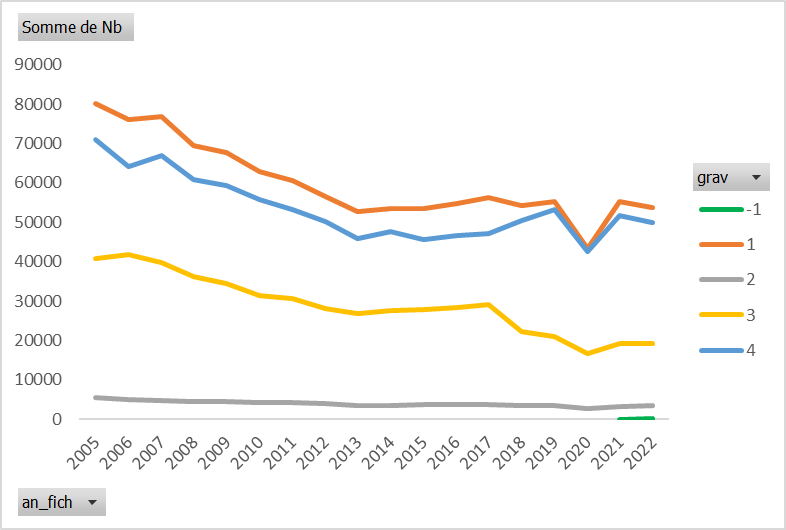
* + - Si cette variable est retenu dans des modélisations futures comprenant des années antérieures à 2018, il faudra vir comment on traite les données de la modalité 4
* **Variable grav**

Cette variable indique la gravité des blessure de l'usager selon les modalités suivantes :

* + 1 – Indemne
  + 2 – Tué
  + 3 – Blessé hospitalisé
  + 4 – Blessé léger

La répartition de cette variable dans le temps montre une variable parfaitement rerenseigné dans le temps

En 2021 et 2022, il y a des modalités « -1 » qui pourrait signifier « non renseigné / non indiqué » mais cela reste négligeable (<1%)



On s’aperçoit que dans la grande majorité des cas, l’usager est indemne.

Il y a peu de tué mais le nombre reste constant alors que les autres modalités diminuent fortement avec le temps

La courbe des blessés hospitalisés et celle des blessés légers suivent une évolution inverse à partir de 2018 (hors 2020) : les blessés hospitalisés tendent à diminuer tandis que les blessés légers tendent à augmenter

Observations / points à éclaircir :

* + - La détermination bléssés léger est subjective, comment est-elle déterminé ?
    - Y a-t-il eu une requalification des type de bléssés ?
    - Que fait-on des « -1 » ?

Glossaire