Haladó adatelemzési módszerek - Accidents in France

Anna Gergály, Júlia Jankó, Péter Mészáros

May 2024

Contents

1 Introduction

The dataset we selected for our project details the accidents that have happened on the roads of France from 2005 to 2016. The dataset contains in-depth information about the vehicles, people and road involved in the accident as well as the weather and other driving conditions. Our goal with this project is to gain insight into what factors play a significant role in roadside accidents and their severity; aiming to be able to predict accident chances and consequences.

We downloaded the dataset from Kaggle where it has been already aggregated for the years 2005-2016, however the data is originally from an official french government website, where it is stored by year. The data is organized into 5 .csv files, one of which contains holiday dates, the other four containing data pertaining to the accidents: characteristics, places, users and vehicles.

You can find our repository here: `https://github.com/annagergaly/french-accidents`

2 Description of files

2.1 Places

Contains data about the place the accident took place, the road quality and other characteristics. There is a lot categorical data describing road type, condition, accident situation, road curvature, traffic type, terrain and infrastructure. These are fairly well filled out and have only a few category options; they are very well usable.

There are a few unknown fields in this file, dealing with a PR value and the french road numbering system, which we do not have further information about.

Out of the remaining fields only three can be considered 'continuous', these all deal with width of the road in some way: number of lanes, width in meters, central width. Sadly these columns have a lot of bad data: values above even 90 for number of lanes and 800 meter wide roads, along with a large number of zero values.

2.2 Characteristics

Contains the characteristics of the accident. A lot of categorical columns can be found in this .csv file, represented in numerical coding. This file also holds information about the the time of the accident, in separate columns for day, month and year.

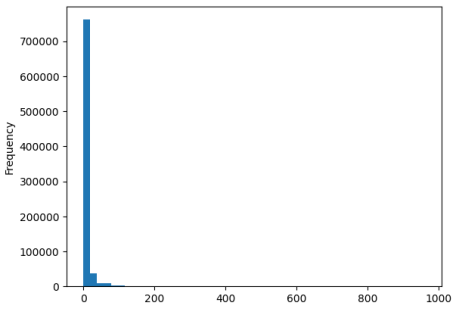
There are a few problematic columns in this file, mostly dealing with spacial data. Address is very specific, down to the street name and as such hard to extrapolate from. There are also problems with missing data, both for the address and the long-lat columns have about 230 000 rows filled out correctly out of the 900 000.

2.3 Vehicle

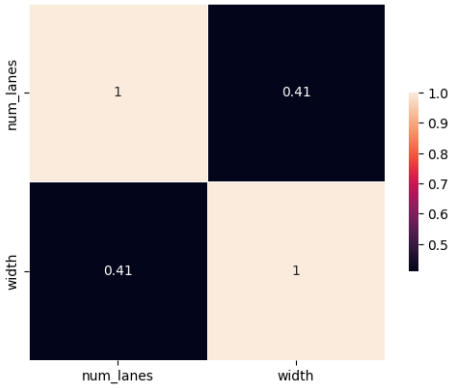
Contains the properties of the vehicles. Flow direction and Occupants columns have a very high 92%-99% null ratio. Flow direction's only gives information about the numbering of the postal address numbers. Occupants is only filled out when public transportation is involved in the accident.

2.4 Users

Contains information about the users (people involved in the accident). Birth year, user type (e.g. driver, passenger, pedestrian), severity are essential information and are almost never missing.



(a) The distribution of central width



(b) Correlation of width and number of lanes

Figure 1: Width and central width

Severity could serve as a main goal of the data analysis, extremities (death and no injury) are the most common. Birth may be useful when considering the driver. There is additional information about pedestrians, their location, action and whether they were accompanied at the time of the accident. Safety equipment is also noted and when used together with the severity of the accident, could reveal useful insights. The type of trip is missing in 29% of cases. Seat in the vehicle is also described, but between 2004 and 2008 it was not recorded.

3 Data preparation and visualisation

Because many variables are categorical, we could not use correlation to measure their pairwise association, instead, we used Cramer's V for this purpose.

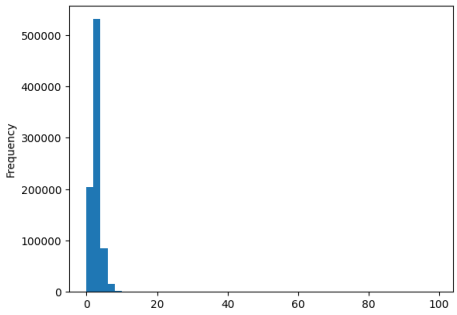
3.1 Places

We have removed quite a few columns while preparing this file. The road numbering system is very opaque and seems unusable for both that reason and because the available data is fairly sparse: v1 and v2 are missing for more than half of rows and road number is 0 for more than half of them.

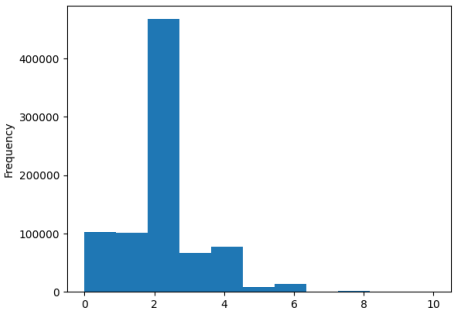
We also removed the columns relating to a so-called 'pr value' (pr and pr distance) as the site contained no information regarding the meaning of this field and a large percentage of the data was missing. There were several columns concerning the width of the road (number of lanes, width, central width) of which the latter two were removed.

Central width had an overwhelming majority of 0 values1a, while width contained very outlandish values ranging from -81 to 999 meters and had a fairly high correlation with number of lanes1b. The number of lanes was mostly correct, but contained way too large values: up to 99, so these were removed.1

Missing values in each column have been filled with either the other/not applicable value (0 in most cases) or with the most common/most likely value for the field (2 lanes for traffic, normal road conditions).



(a) The distribution of lane numbers before data cleansing



(b) ...and after

Figure 2: Data cleansing on number of lanes

The values in this file are almost exclusively categorical and so the used correlation metric was Cramer's V (Figure 3).

3.2 Characteristics

The lighting column values were renumbered from 1 to 5 in a way that now 1 means there was more lighting and 5 means no lighting instead of a random numbering of the categories. The address column contained a lot of missing values and it was sadly not really useful. The most common street address was "AUTOROUTE A1" with 2816 rows which was very little out of the 839.985 rows so we removed it.

Longitude, latitude and GPS columns were removed because department and municipality columns provided enough information to visualise the data and not too detailed for later work. Most other columns contained very few null values and these were filled with the most common values.

There were quite a few correlations in the characteristics file (Figure 4). The most obvious is the holiday column and the date. Holiday is an added column based on the holiday.csv file. This column has a 1 if it's a holiday and a 0 if it's not so it is understandable why it correlates with the time.

Another strong correlation is the Time-Lighting columns. While the time contains the time of day, lighting contains if it's a full day, twilight or dawn, night with or without public lighting. With this information it makes sense that these correlate so notably. Localisation describes if accident took place in more urban or rural areas, so it is expected that it correlates with the value of lighting, because public lighting is more common in built-in areas.

3.3 Vehicle

This data contained very few missing values. For fix obstacles and mobile obstacles columns, null values were filled with the majority value (0) also signaling that there was likely no obstacle if the field was not filled out. For the shock and maneuver columns they were also filled with the most common value, since the few hundred missing values should not be significant compared to the 1.4 million rows of complete data. The flow direction column had 1.3 million missing values, therefore it was removed.

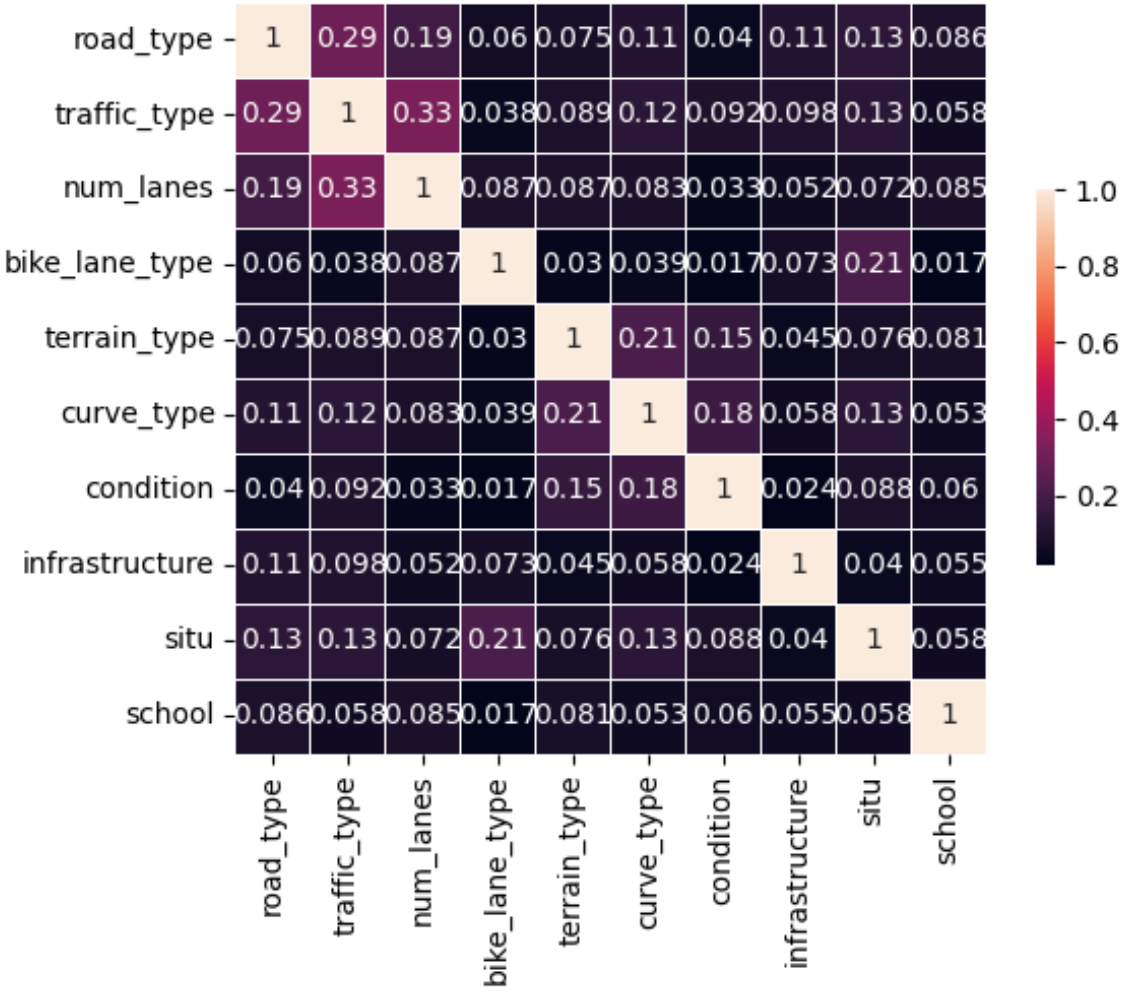


Figure 3: Correlation via Cramer's V in the fields of the places file

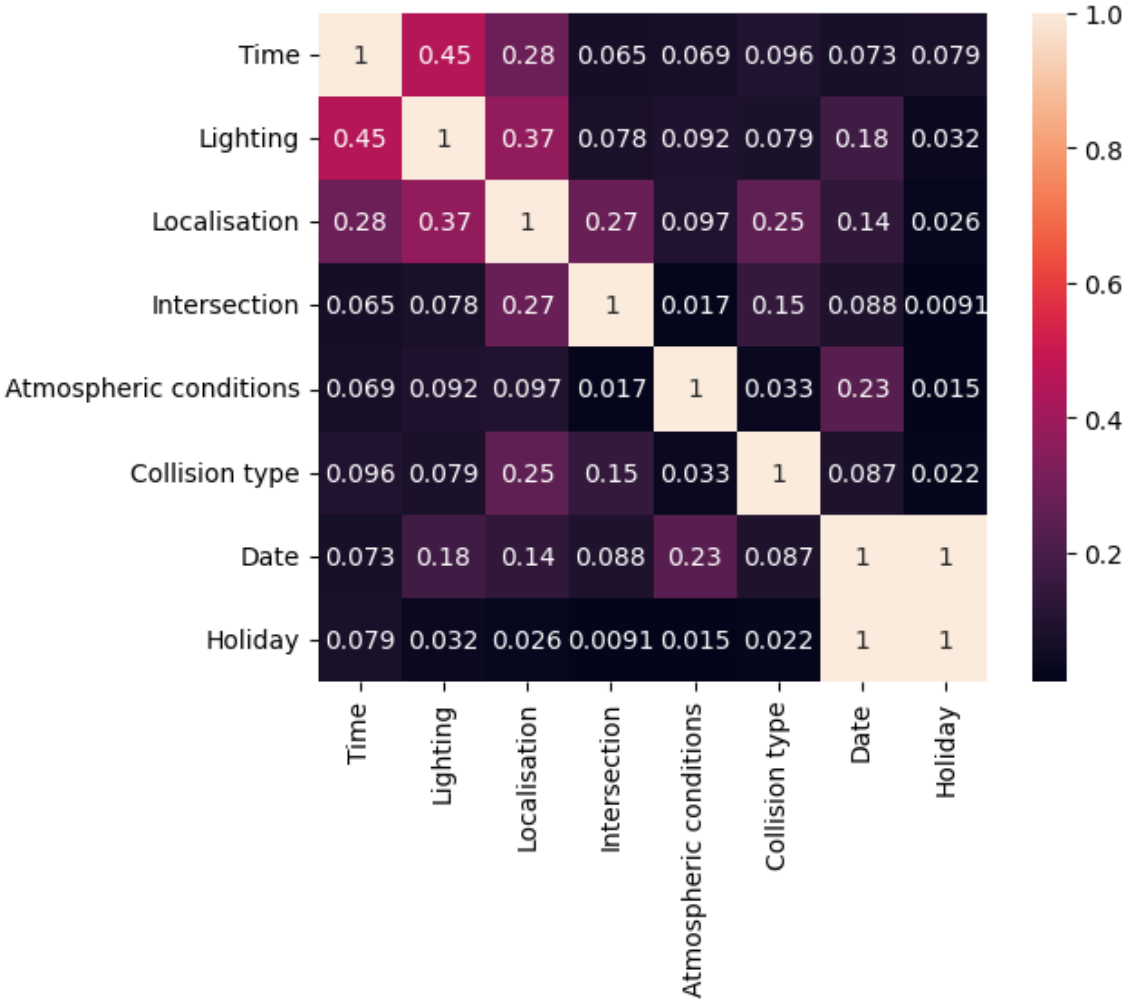


Figure 4: Correlation via Cramer's V in the fields of the characteristics file

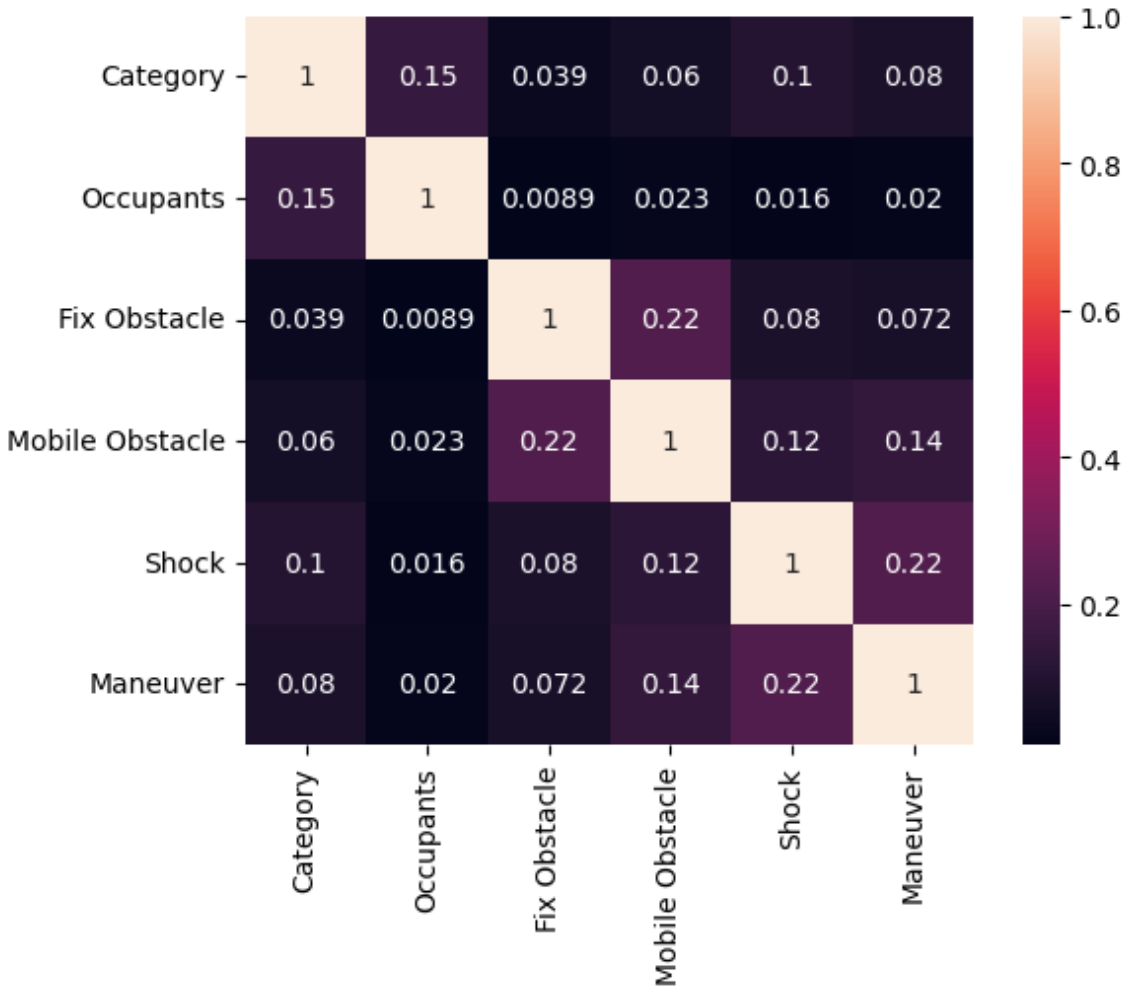
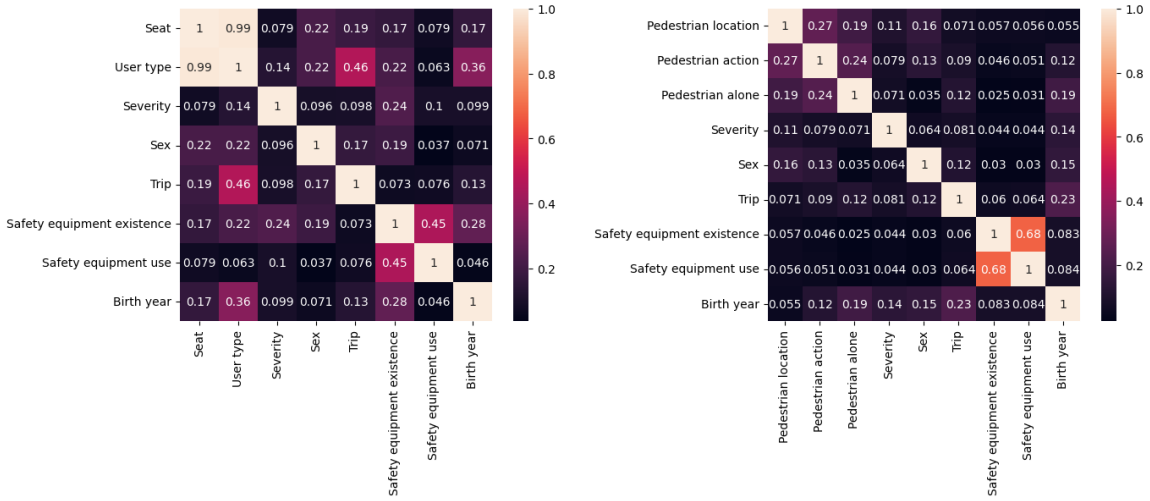


Figure 5: Correlation via Cramer's V in the fields of the vehicles file

When investigating the correlation of the vehicle's properties, there weren't any strong correlations (Figure 5). Only a mild correlation between fix obstacle and mobile obstacle, probably because it is uncommon to have both in an accident. The other correlation is between maneuver and shock, which might be because maneuver could highly influence where the car was damaged.

3.4 Users

In this case, not much data was missing, most columns were almost completely filled out. Seat was missing in four years (between 2005 and 2008), in that case it was filled with 0 to indicate that it is unknown. User type, severity and sex were not missing. Trip was not given in just a couple of cases (about 300). Severity variable was reordered to become a scale which reflects the severity of the accident (1 meaning no harm, 4 meaning death). Safety equipment was not given in about 2% of cases, it was filled with the unknown category. Safety equipment consists of two digits, the first



(a) Not pedestrians (i.e. drivers and passengers)

(b) Pedestrians

Figure 6: Association of the user columns depending on user type

one referring to the existence of a safety equipment (and type of safety equipment), the second one referring to whether it was used or not (or could not be determined), these were separated into two columns.

The association of the variables can be seen on Figure 6. Separating pedestrians and non-pedestrians (i.e. drivers and passengers) is necessary, because pedestrians have additional variables (location, action and whether they were alone) and some variables are not applicable to them (e.g. seat in the vehicle). Another reason for making the distinction, is that an accident can have different outcomes whether one is in a vehicle (and therefore 'protected' in a sense) or not.

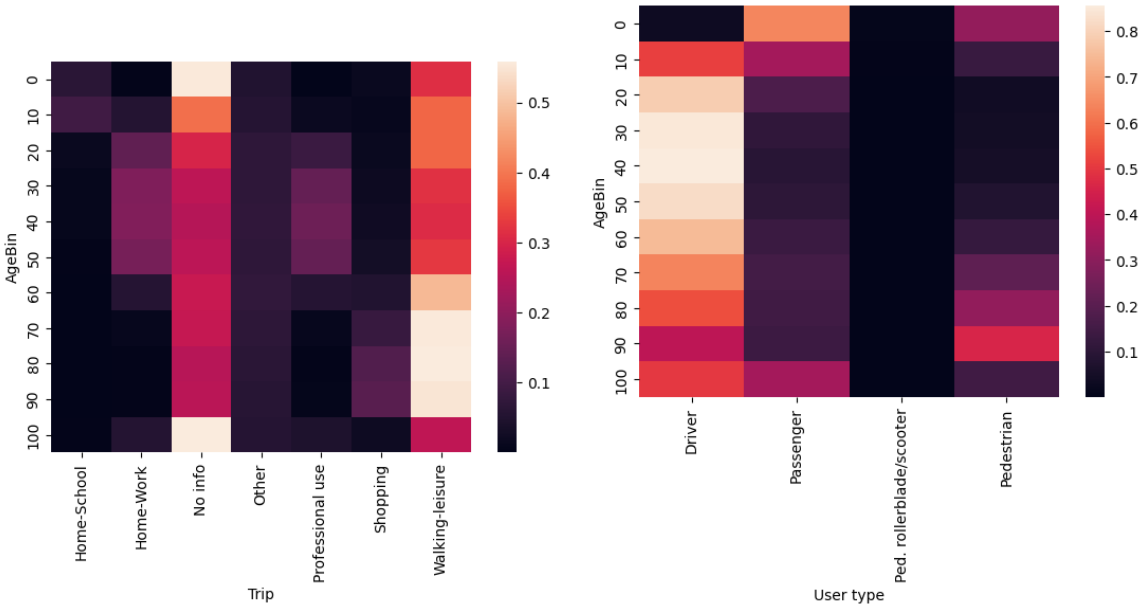
The seat and user type have a high association because in a sense seat further refines the type of user, more precisely the seat of the passenger. In more than 99.9% of cases the driver was in the front left seat (which is expected, because people drive on the right in France). High association is also between safety equipment existence and use, but that could also be due to the missing values (i.e. both variables being 0).

Figure 7 shows the distribution of trip and user types in age intervals with a 10 year step. Younger and very old people tend to be passengers, 30-50 year old people are mostly drivers. The most common trip type was "Walking-leisure", although in many cases the data is not known.

3.5 Merging the dataset

To be able to effectively train a model, we concluded that we should merge our data into one big dataset. In order to do this we have to make as many rows as many vehicles took part in the accident, because each vehicle contains at least one person the vehicle rows all had to be expanded.

This process yielded 1.876.005 rows; each row contains the characteristics of the accident they took part in, the properties of the place and vehicle they were in and the person's attributes.



(a) Trip

(b) User type

Figure 7: Trip and user type category popularities in age intervals. AgeBin refers to age intervals by 10 years, e.g. 0-10, 10-20, etc.

3.6 Visualization on map

We have made two visualizations on a map, using the department codes of the injuries, see Figure 8. (Departments refer to geographical regions/administrative units). Some regions around Paris had no data, but that could be due to department changes along the years. The average severity of injuries vary a bit by department, areas with fewer accidents had higher average severity.

4 Model construction

4.1 Choosing a model and preparing the data

Our chosen model for working with this data was a Bayesian network. This is beneficial for both explainability and querying, which are important in our use-case. The goal is to train a model that can give insight into causality and the interconnectedness of the factors involved in predicting accident outcomes and we want to be able to query the model with a limited number of these factors available.

To make our data compatible with such a network first we deleted the columns used only as database keys and we worked on creating categorical columns with hopefully a low cardinality. To achieve this we quantized our continuous variables, by for example creating age brackets from the date of the accident and the birthyear of the involved persons. Most of the data was already categorical, but we still needed to make changes in some cases because of high cardinality. One of these cases was the department (administrative region) value, where there were in total 101 different values, which needed to be grouped together to create larger geographical units.

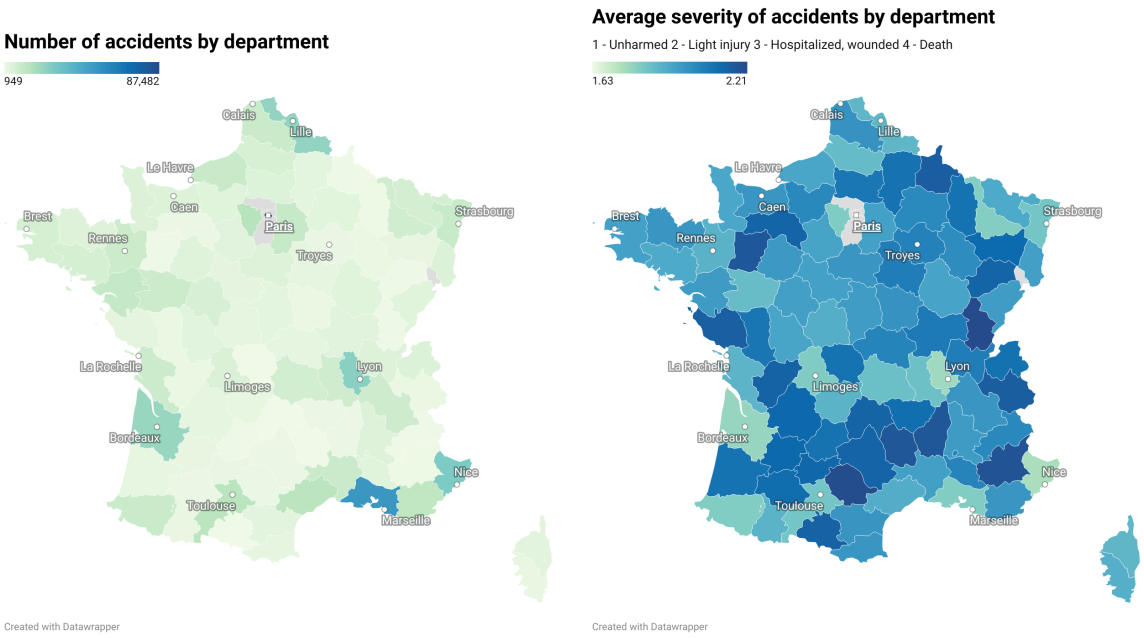


Figure 8: Visualizations by departments.

This way we ended up with around 35 columns for each datapoint, each with under 10 different values, this was however still subject to change while actually fitting our model, as described in the next chapter.

4.2 Working with Bayesian Networks and pgmpy

To create and fit a Bayesian network from just the available datapoints is a very hard task, so we decided to use pgmpy python library which has a couple algorithms implemented for each related task.

To get to a sampleable complete network from the data, we needed to run both structure learning and parameter estimation. Structure learning estimates the shape of the network's graph. Initially we planned to use the min-max hill climb algorithm for this task, but with the large amount of columns in our data, the runtime of this algorithm seemed unfeasable. To deal with this problem we switched to a hill climb algorithm and restricted the maximum in-degree.

To estimate the parameters in the resulting network we used the maximum likelihood estimator included in the library. This process was much faster than the structure learning phase and did not impact our runtime significantly.

The structure of the final model can be seen on Figure 9, which represent the "quasi-causal relations".

4.3 Inference method

For using the resulting model for inference we first tried to use Gibbs-sampling, but this failed because of high memory demands. We then experimented with exact inference using variable elimination,

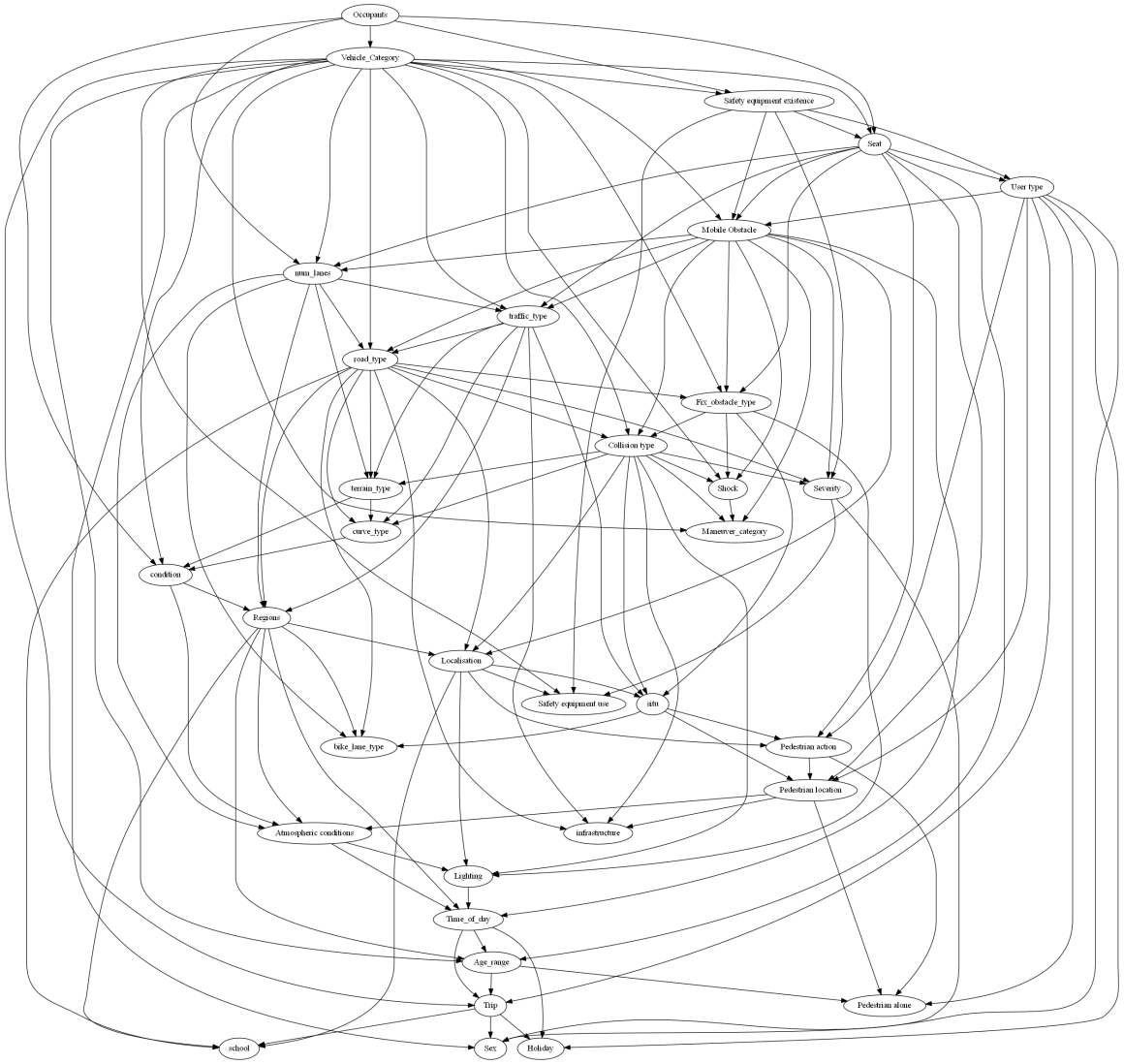


Figure 9: Visualization of the structure of the learned model, with a restriction of each node having at most 4 incoming edges.

which proved to be quite successful on the higher memory capacity machines (16-32 GB) that were available to us.

Providing more evidence generally makes these queries easier and faster to run and even with this inference method single evidence queries are not always possible, but it generally works very well for answering interesting questions about the dataset.

5 Answering questions

Our approach was to ask questions and sort of translate them for the model through our available evidences. We then examine the results to see what could be the cause of the result and how it aligns with previous knowledge, biases and assumptions.

5.1 Severity without evidences

We created a query just to see what are the computed chances for the different severity classes of injuries without any evidence supplied, so that we are able to refer to these as a baseline whenever discussing the distribution of severity in more specific cases. The results are the following: it is most likely they will be unharmed (41%), a high chance of light injury (36%), a relatively high chance of being hospitalized (21%) and a low chance of death (3%).

5.2 Analysis of bicycle accidents

These questions relate to bicycle accidents. In Île-de-France region a cyclist, who does not use safety equipment has a 3.8% probability of dying (compared to 2.7% generally). With the use of safety equipment, this reduces to 3.2%, but they are more likely to be hospitalized.

If someone dies in a biking accident (with or without safety equipment), they were most likely in the Île-de-France region. This is generally true, of all biking accidents. A reason for this could be that, Paris is in that region, which is regarded as a bike friendly city, and there probably there is a deep culture of cycling and well-built infrastructure.

During the night, cyclist are twice as likely to die, than in proper light conditions. With the use of safety equipment, the probability of dying is less (6% as opposed to 7.7%).

5.3 Motorbike users

The query is about the age range, if the driven vehicle is a motorcycle. There is about a 50% probability of being in the 25-45 group.

5.4 Gender and severity

These questions revolve around how severe the injuries where of men and women in the same conditions. Driving in daylight in the city, men are more likely to have no injuries than women, women are more likely to have light injuries and a bit more likely to be hospitalized, but men are more likely to die. Men are more likely to be drivers (76% vs. 60%) when in an accident in these conditions, and women are more likely to be pedestrians or passengers. In rural areas, both are more likely to be injured.

5.5 Age and severity

The queries in this case are asking the age range of the persons given a few evidences. If a person is hospitalized and a driver of a car, it is likely to be in the 25-45 age group. If a driver uses safety equipment, the probability of each age group corresponds to the distribution of that age group.

5.6 Trip reasons

The question was the following: "What are the likely reasons a driver 25-45 years old vs 16-25 started a trip that lead to hospitalization when driving during the day?". The result is that among both groups, leisure trips are the most common. Naturally, home-work trips are more common among 25-45 year olds, and home-school trips are more common among 16-25 year olds. Professional use is understandably more common in the first group.

5.7 Deadliest road conditions

These queries asked the road condition, terrain type and road curve type variables based on the following evidences: the person is a driver aged between 25-45 and is killed in the accident. Normal and wet road conditions had 75% and 17% probability, respectively; flat terrain and slope had 70% and 16%; and lastly, straight road and road curving in either direction had 70% and 22%.

5.8 Severity during the holidays

The question this query revolves around is whether drivers are more likely to be injured worse during the day on a holiday. The results suggest no significant difference in the probabilities regarding the severity of the accident.

5.9 Significance of weather conditions

Here we query about how the weather factors influence the severity of traffic accidents. Our model suggests that during light rain people are less likely to get injured (0.14% more likely to be unscathed). When taking light rain conditions, our model shows that death and hospitalization is less likely and people are more prone to light injuries (1.2% more lightly injured people). Heavy rain causes higher death (0.26%) and hospitalization (0.77%) rates. This might mean that during rain drivers are more cautious but light rain doesn't really cause more accidents.

6 Comparing our results to previous literature

We have looked into papers regarding causes and correlations relating to traffic accidents, primarily in France but also considering western Europe more broadly, as we believe driving habits are similar in this region. Our goal is to see how our results compare to previous literature, especially ones based not only in statistical results, but also expert opinion from related fields.

6.1 A possible missing link

A lot of relevant literature and in particular [5] from our more closely surveyed literature mentions the effects of drugs and in particular alcohol on traffic accidents. The authors of this paper worked with an older police dataset from September 1995 to December 1999 on accidents involving less than

3 vehicles (alcohol test results were available for 78.7% of drivers) and analyzed the data descriptively and utilized logistic regression.

They found alcohol consumption to be an immense factor (even the main factor) in both the occurrence and the causalities of an accident. In deadly accidents the driver has consumed alcohol over the legal limit in 31.5% of cases. Single vehicle accidents stood out as especially influenced by alcohol, in this case this factor overwhelmingly dominated all other considered (age of driver, meteorological conditions, time of day).

In our dataset there was no data regarding any type of substance use or blood alcohol levels, which is quite unfortunate and even surprising considering the police origin of the data. This removed an important factor that we could have analyzed.

The closest we can get to this feature is trying to look at situation where it is most likely that a participant is intoxicated. In the paper they have found that at night on the weekend in single-vehicle fatal accidents a staggering 71.2% of drivers were under the influence.

6.2 Analysis of road users

We attempted to answer a lot of questions related to different aspects of road users and in our review we found [2] to be a valuable resource for comparison. The paper focuses on the 2007-2008 timeslot (also contained in our data) and other than crash data, also utilizes exposure data extracted from the french national household travel survey.

This extra layer of information can show us further insight into certain figures we extracted from our data or even contrast our findings. The researchers assessed the number of trips, distance traveled and time spent traveling by road user type, age and sex.

People from the 17–20 and 21–29 age groups and those aged 70 and over were the most at risk groups according to the paper, with the 17-20 bracket having an outstanding 142.7 per 1 million inhabitant fatality rate. This is in line with our findings regarding age.

In [3], it has been found that accident severity increased with age, car drivers accounted for the majority of casualties.

The risk of being killed was 20 to 32 times higher for motorized two-wheeler users than for car occupants, which shows motorbike usage to be extremely dangerous. For cyclists, the risk of being killed, both on the basis of time spent traveling and the number of trips was about 1.5 times higher than for car occupants. These travel methods are generally cheaper than car usage and are more attractive to young people, especially in lower income brackets, which might be part of the age related differences.

Risk for pedestrians compared to car occupants was similar according to time spent traveling, lower according to the number of trips and higher according to the distance traveled. These differences come down to how walking trips tend to be shorter but much slower.

According to [2] male travelers had a higher fatality rate than females, by a factor of between 2 and 3. This is starkly different than what we have found, where sex had very little sway in the results of our queries. The reason for this is most likely our lack of data about accident-free usage and might suggest that men and women have similar outcomes in accidents, but on average men are involved in much more.

However a different study [4] about occupational disparities in accidents suggests that women generally take part in more traffic accidents than men (the ration is 0.88 to 1). The study focused on north-eastern France and the time period from 2006 to 2008 which overlaps with our dataset. This paired with the previous article which is also in this time period suggests that men have higher fatality rates even though they take part in less traffic accidents than women.

6.3 Bicycle usage

Bicycle accidents in Belgium were examined in [6]. The inference we conducted in regards for cycling was covering the accidents, which happened in the Île-de-France region of France, which contains Paris and its surroundings. Its safe to assume that the comparable region to that in Belgium is Brussels and its surroundings.

In [6], it has been found that the risk of cyclists getting hospitalized because of an accident decreases as the proportion of cyclists increases. We found that if someone dies in a bike accident, it is most likely to have happened in the Île-de-France region, which is somewhat contradicting to the results in the Belgian cycling paper, but they examined the hospitalization (meaning that the cyclist was injured so badly that they required hospitalization), not death, which only is a portion of the latter.

6.4 Weather conditions

We wanted to investigate the impact of weather conditions further and found this [1] article quite useful for that. This study focuses on injury accidents in France, the Netherlands and the Athens region. This research is based on police records from 1975-2000 in France. While this does not directly overlaps with our dataset, it still can be relevant.

They concluded that the rainfall and temperature is positively correlated with the number of injury accidents. 100 mm of additional rainfall during a month increases the number of injury accidents in that month by 0.2–0.3%. As for the temperature, they found that 1 °C of additional average temperature during a month increases the number of injury accidents in that month by 1–2%. They also conducted research on the impact of frost but that seemed to negatively correlate with the number of injury accidents. 1 additional day of frost during a month decreases the number of injury accidents in that month by 0.3–0.6%.

While we don't have detailed data about the weather conditions, merging heavy and light rain probabilities compared to normal conditions, there is no significant change in injury occurrence. This contradicts with the findings of [1]. The cause might be the differences of the time period or road quality. The lack of detailed weather data might also be the cause of these differences. Sadly we don't have data on temperature and frost conditions.

6.5 Occupational hazards, work related accidents

In [3], the work-related and non-work-related accidents were examined for accidents between 1997 and 2006, divided into two periods, 1997-2000 and 2003-2006. The latter has a 2 year overlap with our dataset, which starts from 2005. Since then, many factors could affect a change in road safety (e.g. a change in driving culture, the installation of automatic radars). In the study, they only included the drivers and focused on 14-64 year olds. They considered 4 types of trips: commuting (to or from work), journey while at work, other and unknown.

Based on the data provided in [3], the casualties (regardless of severity) in the 25-45 age group in a work related trip ("professional use" is the closest trip type in our dataset) in the 2003-2006 year range take up 12.3%. Based on our model, given that the accident is during the day, the driver is in the specified age range and hospitalized, being a professional trip has a 14.5% probability, which is comparable. Important to note that this does not include deaths and light injuries. With the same same conditions as above, commute takes up 20.3%, based on our model, it is about 24%. The probabilities given by our model are not too far off from these findings.

7 Final analysis

7.1 Possible use-cases

Since the model is capable of predicting the likelihood of different levels of injuries, it could be suited for different kinds of risk-management applications.

This could include for example first responders: a model could supply them with the probability of different types of injuries in different types vehicles in the given region and the current weather conditions. This, especially paired with information about exposure could help with distributing resources and getting ready for events.

It could also be useful for the average person getting ready for a trip: this type of information could be included in a route planning applications, possibly supplying a 'safest' path to take to motorists, pedestrians and cyclists.

Other than these, this sort of more intuitive representation of this crash data can also be helpful for decision makers, politicians or urban planners: it could help bring large infrastructural gaps and particularly dangerous areas or traffic solutions.

7.2 Legal and ethical concerns

In a lot of machine learning applications the main legal and ethical concerns stem from the data acquisition: gathering data about individuals without consent and violating data protection laws. In our case, these concerns luckily do not apply: all of our data is from public, government published sources and does not contain sensitive information on the individuals involved in the accidents, nor does it make any person identifiable.

The use of the finished model is also a very important part of making sure everything is in accordance with legal and ethical standards. Since the used Bayesian network is very explainable (as opposed to a more black box-like architecture, like a neural network) and all results can be tied to statistical properties of the input data.

This means that while the model can be a very useful aid in gaining new insight from the raw datapoints, bringing a more thorough statistical understanding, it does not make decisions 'on it's own' and helps maintain transparency in the decision making process.

References

[1] Ruth Bergel-Hayat et al. "Explaining the road accident risk: Weather effects". In: *Accident Analysis & Prevention* 60 (2013), pp. 456–465. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2013.03.006. URL: https://www.sciencedirect.com/science/article/pii/S0001457513000948.

[2] Liacine Bouaoun, Mohamed Mouloud Haddak, and Emmanuelle Amoros. "Road crash fatality rates in France: A comparison of road user types, taking account of travel practices". In: *Accident Analysis & Prevention* 75 (2015), pp. 217–225. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2014.10.025. URL: https://www.sciencedirect.com/science/article/pii/S0001457514003224.

[3] Barbara Charbotel, Jean Louis Martin, and Mireille Chiron. "Work-related versus non-work-related road accidents, developments in the last decade in France". In: *Accident Analysis & Prevention* 42.2 (2010), pp. 604–611. ISSN: 0001-4575. DOI: https://doi.org/10.1016/j.aap.2009.10.006. URL: https://www.sciencedirect.com/science/article/pii/S000145750900270X.

[4] M. Khlat et al. “Occupational disparities in accidents and roles of lifestyle factors and disabilities: a population-based study in north-eastern France”. In: *Public Health* 122.8 (2008), pp. 771–783. ISSN: 0033-3506. DOI: https://doi.org/10.1016/j.puhe.2007.09.012. URL: https://www.sciencedirect.com/science/article/pii/S0033350607003204.

[5] Michel Reynaud et al. “Alcohol is the Main Factor in Excess Traffic Accident Fatalities in France”. In: *Alcoholism: Clinical and Experimental Research* 26.12 (2002), pp. 1833–1839. DOI: https://doi.org/10.1111/j.1530-0277.2002.tb02490.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1530-0277.2002.tb02490.x.

[6] Grégory Vandenbulcke et al. “Mapping bicycle use and the risk of accidents for commuters who cycle to work in Belgium”. In: *Transport Policy* 16.2 (2009), pp. 77–87. ISSN: 0967-070X. DOI: https://doi.org/10.1016/j.tranpol.2009.03.004. URL: https://www.sciencedirect.com/science/article/pii/S0967070X09000407.