UK Road Traffic Collision

Nonparametric Statistics project

Valeria Iapaolo, Oswaldo Morales,
Riccardo Morandi, Abylaikhan Orynbassar,

12 December 2023

POLITECNICO
MILANO 1863

Dataset

Collision data

For each **collision** we know:

- **Date** and **time**;
- Geographical **location** (latitude and longitude);
- Local authority **district**;
- **Road** type and conditions;
- **Weather** conditions.

Vehicle data

For every **vehicle** involved in each accident we have:

- Vehicle **type** and **propulsion**;
- Vehicle **manoeuvre**;
- Vehicle **age**;
- **Point** of **impact**;
- **Position** in carriageway;
- Age and sex of the **driver**.

Casualty data

For every **casualty** of each vehicle we know:

- Casualty **severity** (slight, serious, fatal);
- **Age** band and **sex**;
- Casualty **class** (driver/rider, passenger, pedestrian, ...)
- **Position** on the road in pedestrian case.

We used **official** data from the UK's **Department of Transport**, we decided to focus on the years from **2005** to **2022**.

Nonparametric Tests and ANOVA

Significance of casualty class on casualty severity

Similarly, to test the whether the Pedestrians, Drivers and Passengers (casualty_**class**) have the same **severity** (Slight, Serious, Fatal) , we performed **permutational** ANOVA**:**

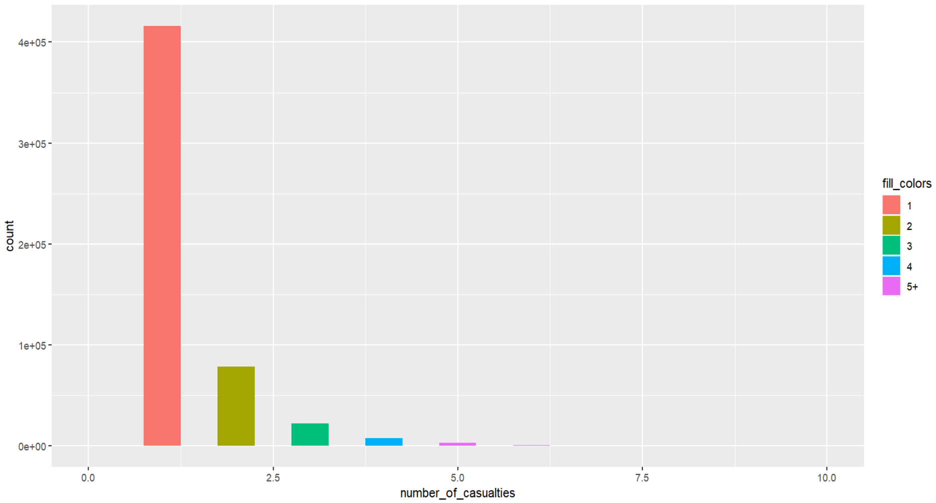
casualty_severity ~ casualty_class

```
                Df Sum Sq Mean Sq F value Pr(>F)    
casualty_class    2     26  12.988   84.97 <2e-16 ***
Residuals      9997   1528   0.153                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- As it can be seen from the summary of the test we can conclude that the casualty class has a **significant effect** on the casualty severity.
- Test is performed only on the **subsample** of the data.

GAM: number of casualties

For each accident we model the number of casualties a **GAM** assuming a **zero inflated Poisson** distribution to take into account the large number of accidents with only **one** casualty.



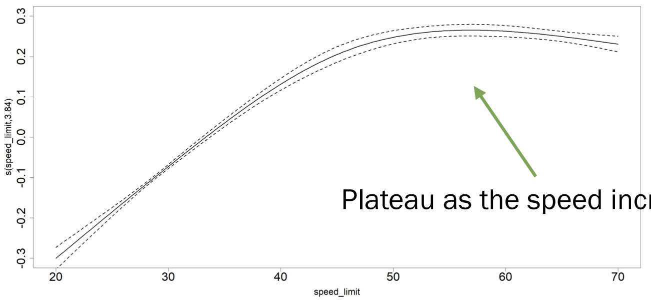
The zero-inflated Poisson (**ziP**) model mixes **two** generating processes.

- The first process generates **zeros;**
- The second process is governed by a **Poisson** distribution that generates counts, some of which may be zero.

$$
\begin{aligned}
ziP(number\ of\ casualties - 1) \sim \\
weekend\ + \\
light\ conditions\ + \\
s(time, bs\ =' \mathrm{cc}') \\
+\, s(number\ of\ vehicles, bs\ = \text{‘cr’}, k\ =\ 6) \\
+\, s(speed\ limit, bs = 'cr', k\ =\ 5)
\end{aligned}
$$

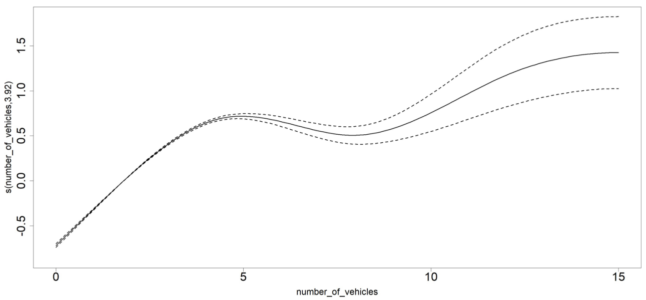
Results

Speed limit



Plateau as the speed increases

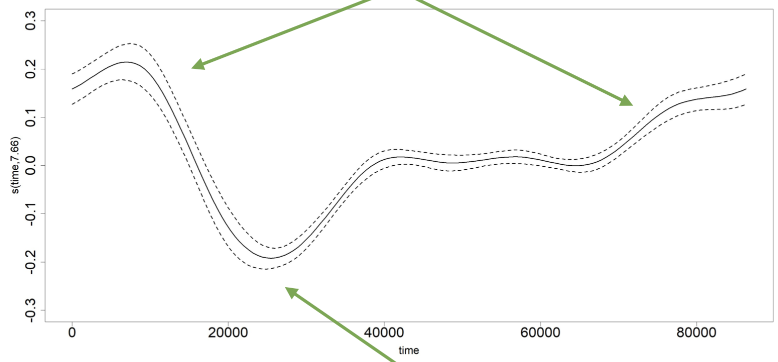
Number of vehicles



Time

The time behaviour is non-trivial, with a clear increase of the number of casualties in the night.

Increase during night-time and early morning hours



Decrease in the day, probably due to the prevalence of solo commuters.

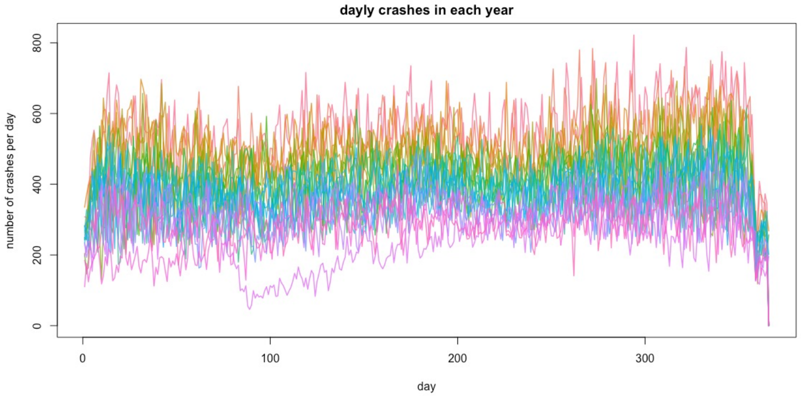
Functional data

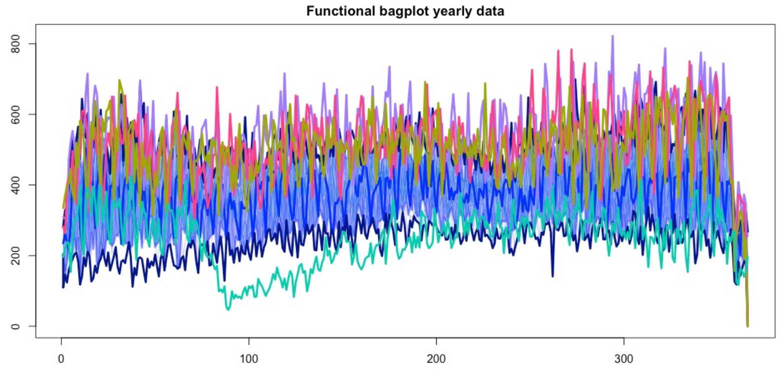
We model the **number** of **crashes** as **functional** data, focusing on **4** different **time horizons**: year, month, week, day.

Yearly data

We found:

- A clear **outlier** in the year 2020, due to the covid pandemic;
- A clear **decreasing trend** as the years progress despite the increase in circulating vehicle.





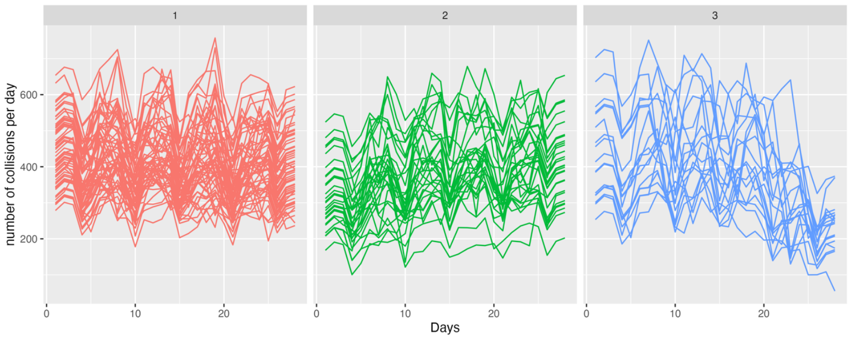
Functional data

Monthly data

When considering monthly data, we decided to **align** the data using **shift warping function** to properly capture the weekly pattern in the data.

Functional clustering using k-means

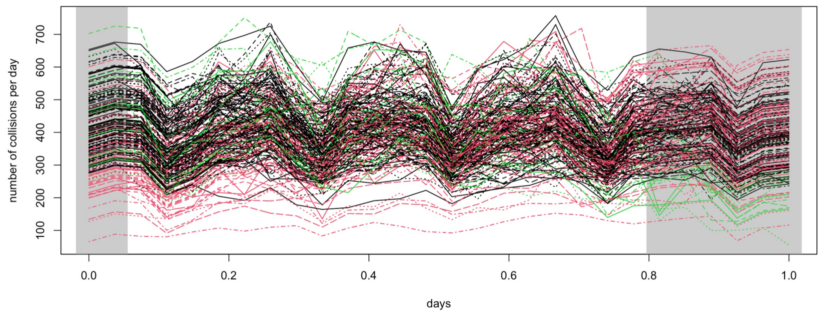
We found **3 clusters**: one containing predominantly the months of **January** and **April**, one containing mostly **December** and the **remaining** months were clustered together.



Permutation tests on the identified clusters

We validated the results using permutation tests:

- **Global** permutational **ANOVA**;
- **Local** permutational **ANOVA** using an **interval-wise** testing procedure.



Functional data

Weekly data

There were two distinct clusters:

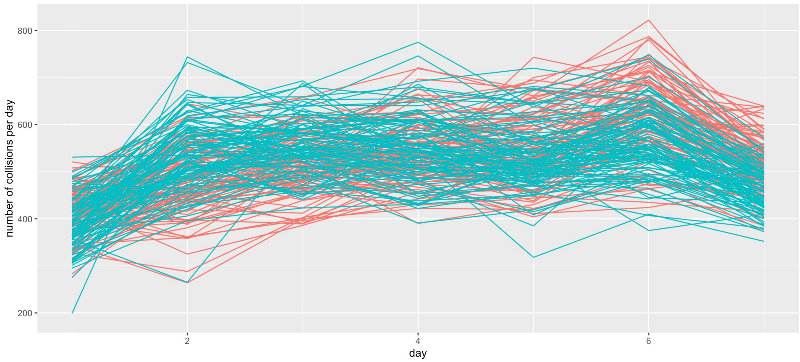
- Regular working weeks;
- Weeks belonging to a holiday period.

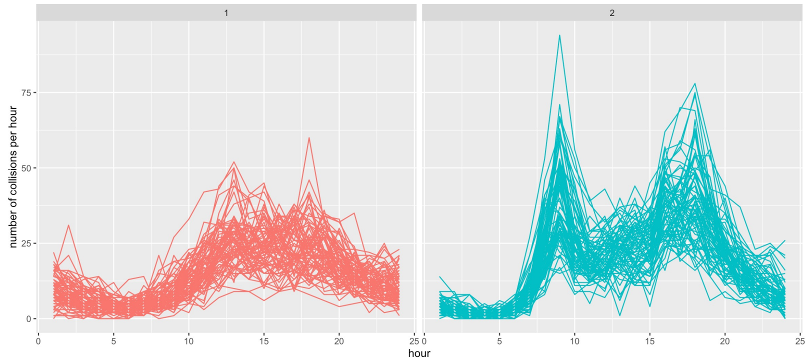
Daily data

There were two distinct clusters:

- Regular working days;
- Weekends and holidays.

The significance of the clusters was validate using a **global permutation** 2 population test for the difference in distribution. From a **local** permutation test the distribution was different **on the whole time span.**





Nonparametric Tests and ANOVA

Significance of latitude and longitude on the number of crashes

In order to test the significance of the geolocation we used a **two-way Permutational ANOVA** with interactions:

```
number_of_crashes ~ binned_longitude +  binned_latitude
          + binned_longitude:binned_latitude
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
binned_longitude	1	290	289.89	119.61	< 2e-16	***
binned_latitude	1	238	238.19	98.28	< 2e-16	***
binned_longitude:binned_latitude	1	147	147.14	60.71	6.98e-15	***
Residuals	17824	43199	2.42			

- **Both** the **longitude, latitude** and their **interaction** have a **significant** impact on the number of crashes. This suggests that geographic location is an important factor in traffic accidents.
- The test is performed only on the **subsample** (30000) of data due to the memory and time constraints.
- **Permutational** ANOVA was used because the data do **not** follow the Gaussian distribution.

Number of daily collisions per district

GAM model with mixed effects

$$\log(mean\ n^{\circ}\ of\ collisions_i) \sim day\ type_i + wind_i + rainfall_i + year_i + f_1(day\ of\ the\ year_i) + b_i$$

Rainfall → Fine, Rain, Snow

Day of the year: 1, … , 365
modelled using a **cubic spline**

Random intercept for the **local authority district** used to capture the difference in the population across the country.

Year: 2005, … , 2019

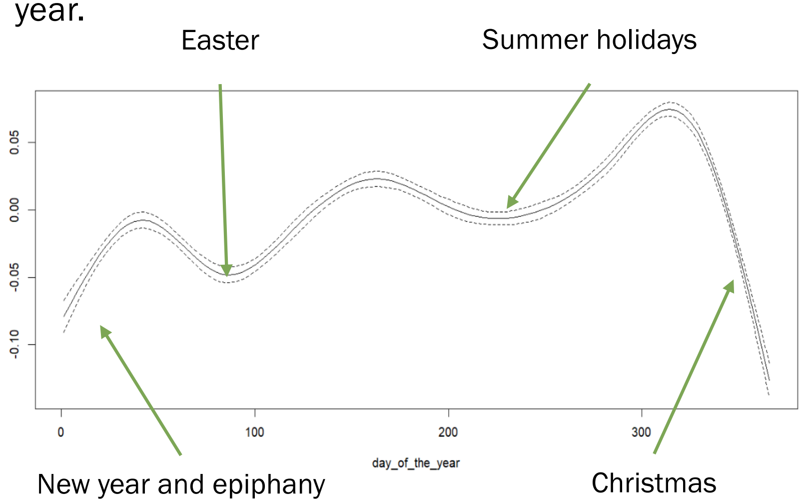
Day type → Weekday, Weekend

Wind → High wind, No high wind

Results

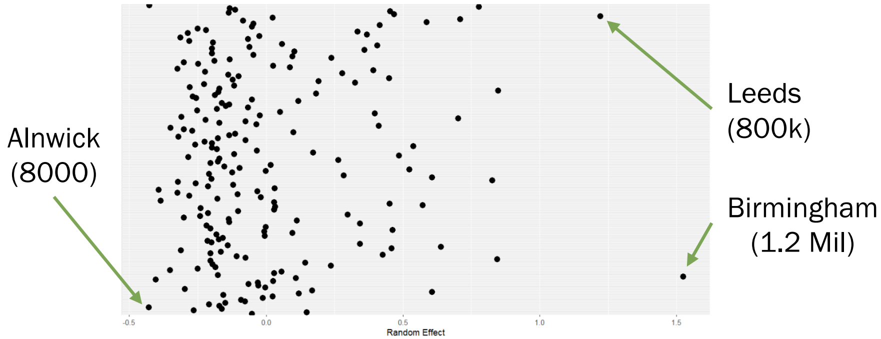
Day of the year

The nonparametric part of the model is able to correctly **capture** the **nonlinear behaviour** of the number of crashes in the **different periods** of the year.



Random effects

The random intercept correctly accounts for the **population differences** of the districts.



Rain fall

In Birmingham the 1st of Feb 2019 the predicted mean number of crashes are:

Rainfall	Mean n° of collisions
Fine	6,78
Rain	6,43
Snow	6,20

Challenges and Next Steps

- Improve the **computational efficiency**, possibly with a smart subsampling, when performing both permutation tests and GAM models. The challenge is that most of the information is **crash specific** and is **lost** when aggregating data.

- **Improve** the GAM models, both in terms of model performance and goodness of fit:
  - When modelling count data, the Poisson assumption is **not** satisfied;
  - The random effects do **not** follow a normal distribution;
  - Perform **permutation tests** to assess the significance of the coefficients.

- Perform **conformal prediction** for both:
  - The **G**eneralize **A**dditive **M**odel approach;
  - The **functional** data approach.

- Further incorporating the **spatial information** in the model via:
  - Adding **latitude** and **longitude** as regressors in a **GAM** model;
  - **Nonparametric spatial** model to estimate the **trend** and the **variogram** *(npsp);*

  The challenge is that most of the data is located around large cities and along the motorway network.