UK Road Traffic Collision

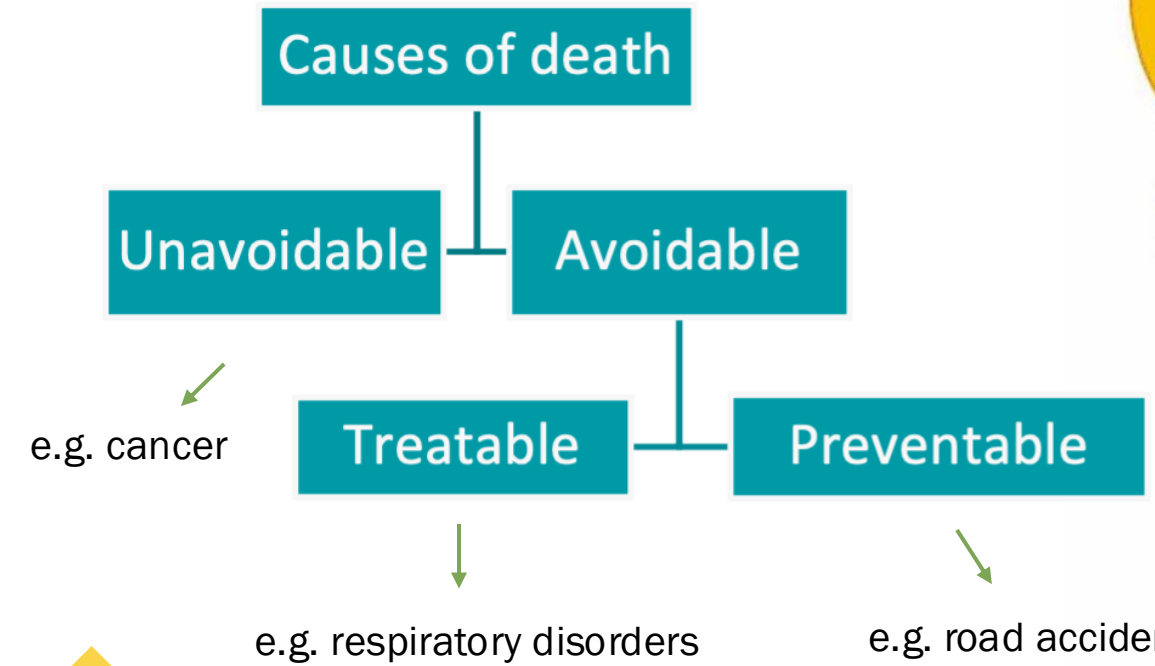
Nonparametric Statistics project

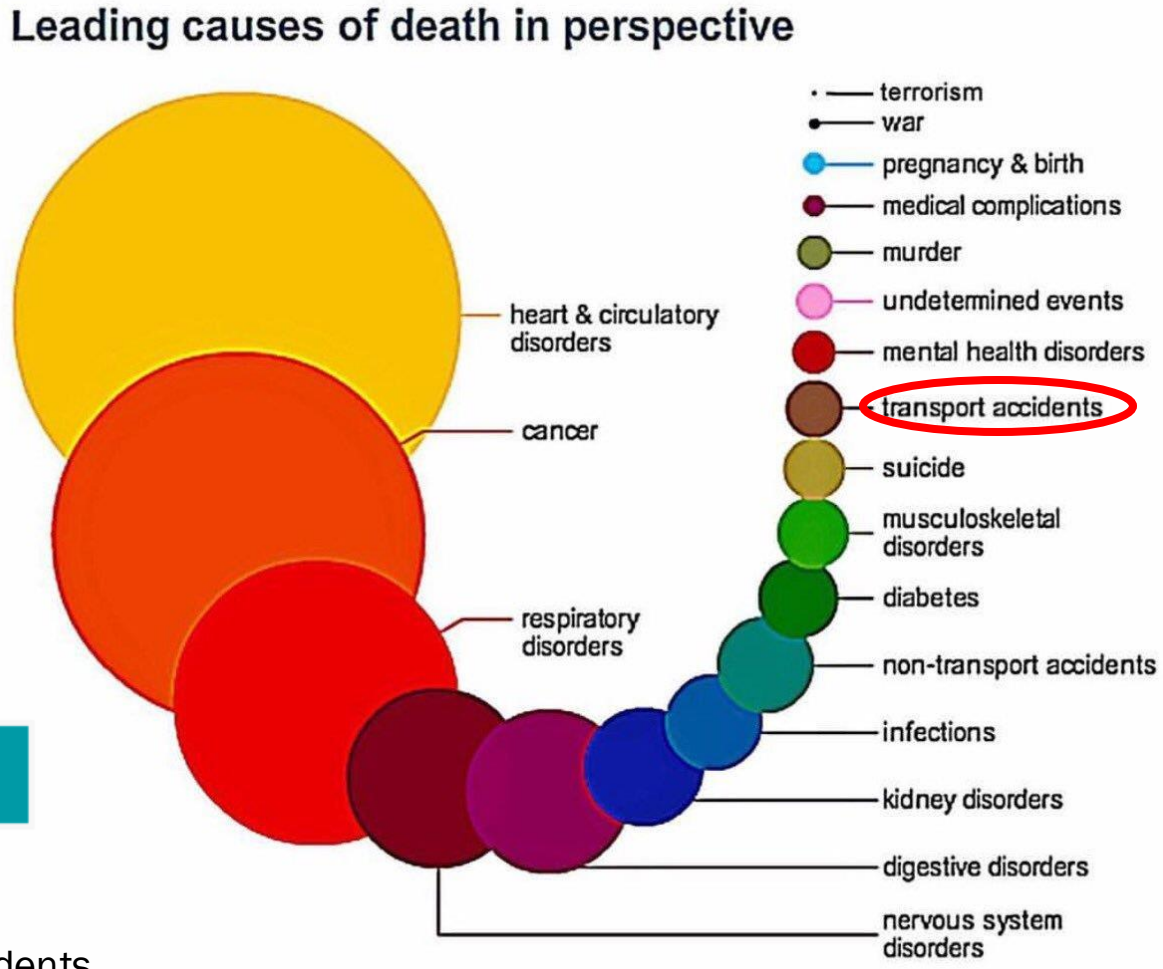
Valeria Iapaolo, Oswaldo Morales,
Riccardo Morandi, Abylaikhan Orynbassar,

16 February 2024

POLITECNICO
MILANO 1863

Causes of death





According to the latest data published by the WHO in 2020, deaths from road accidents is the $\mathbf{30^{th}}$ **cause of death** (0.42% of total deaths).

Stakeholders

Understanding the potential number of accidents that may occur in a specific area and in a specific time slot provides support for the planning of **police** and **first aid** services.

Employee Work Schedule

MONDAY	TUESDAY	WEDNESDAY	THURSDAY	FRIDAY
SHIFT 8 am - 12pm	SHIFT 8am - 4pm	DAY OFF X	SHIFT 8am - 4pm	SHIFT 8am - 4pm

SATURDAY	SUNDAY	NOTES
SHIFT 8am - 4pm	DAY OFF X	1. Give the key to Ann.

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.

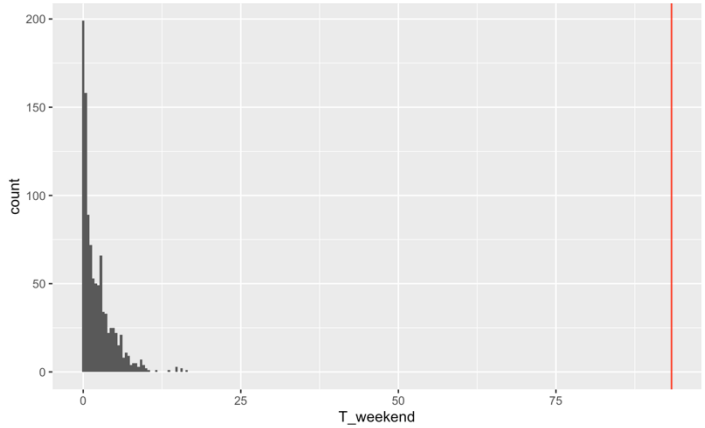
We used **official** data from the UK's **Department of Transport**, we decided to focus on the years from **2005** to **2022**.

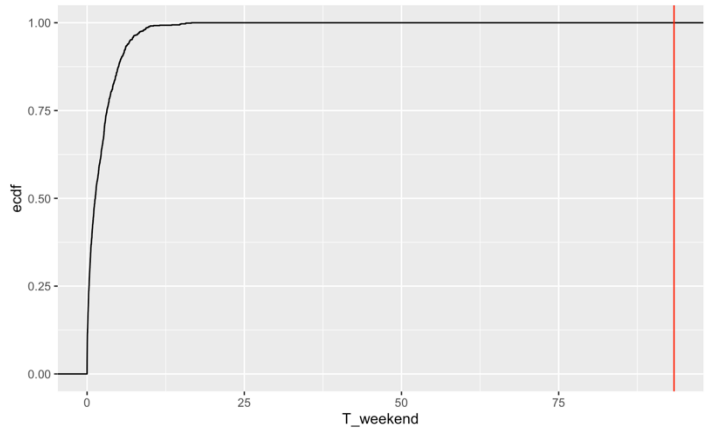
Nonparametric Tests and ANOVA

Significance of number of casualties, light condition and weekend class on casualty severity

```
casualty_severity ~ number_of_casualties + light_conditions + weekend
```

The results of permutational suggest that all the covariates are significant, with p-value of zero.





Accident severity

GAM for the severity of an accident

We employed 2 **GAM** models to analyse the accident **severities**:

1) We started by analysing the **3 categories** of severities, **slight, serious and fatal**, with an ordered categorical family with the identity link function: $g_1(y) = y$.
2) Then we reduced the response to **2** categories: **slight and serious-fatal**. The same semiparametric model for the covariates was maintained, and the link function was changed: $g_2(y) = \log(y)$.

$$g_j(accident\ severity_i) \sim number\ of\ casualties_i + weekend_i + light\ conditions_i + f_1(time_i) + f_2(number\ of\ vehicles_i) + f_3(speed\ limit_i)$$

$$i = 1, \ldots . n, \qquad j = 1, 2, \qquad f_1, f_2, f_3\ are\ cubic\ splines, in\ addition\ f_1 is\ periodic$$

Due to imbalanced of the data towards the class "slight severity“, we kept the second model.

Results

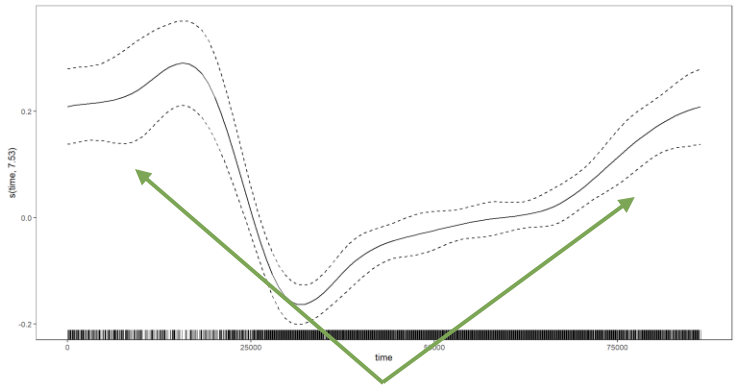
Parametric coefficients

Parametric coefficients	binary model
Intercept	-1.506703
number of casualties	0.207505
weekend	0.072481
Dummy for light conditions: "Darkness - no lights"	0.107128
Dummy for light conditions: "Daylight"	0.002283

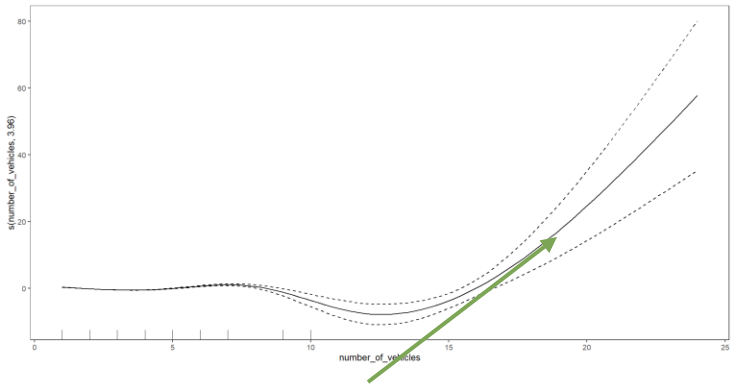
Comments:

- **Positive influence** on the severity of collisions by all the parametric coefficients;
- Coefficient of the dummy variable Daylight is positive, but it was the only coefficient that did not show statistical significance in the parametric test conducted by the model.

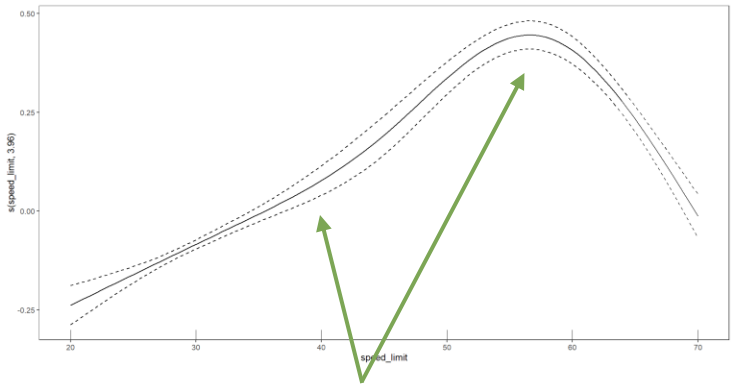
Splines



Increase during night-time and early morning hours



Clearly increasing as the alongside the number of vehicles



speed limit increases severity until ~60mph (further investigations are warranted)

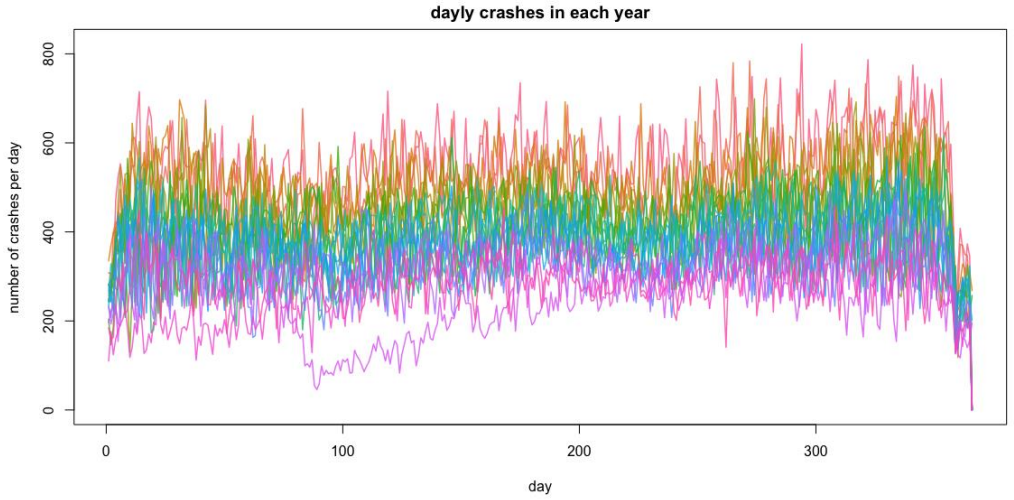
Functional data

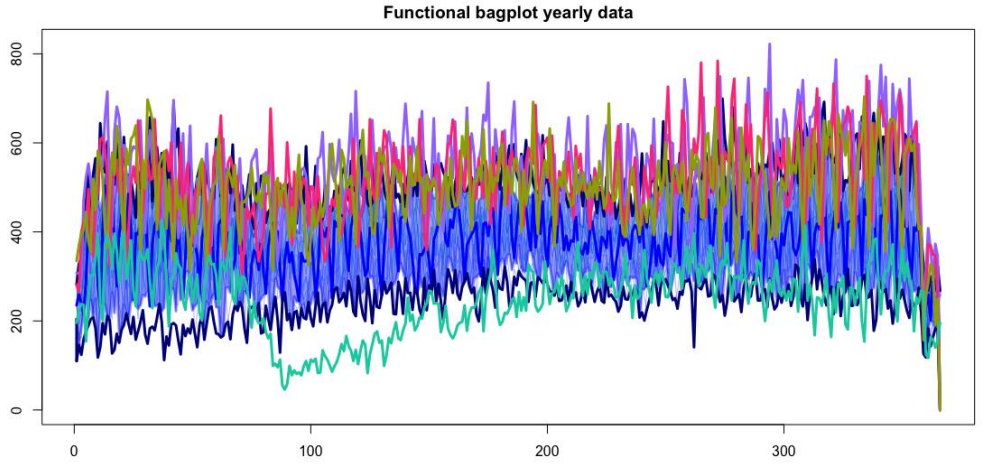
We model the **number** of **crashes** as **functional** data, focusing on **4** different **time horizons**: year, month, week, day.

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.





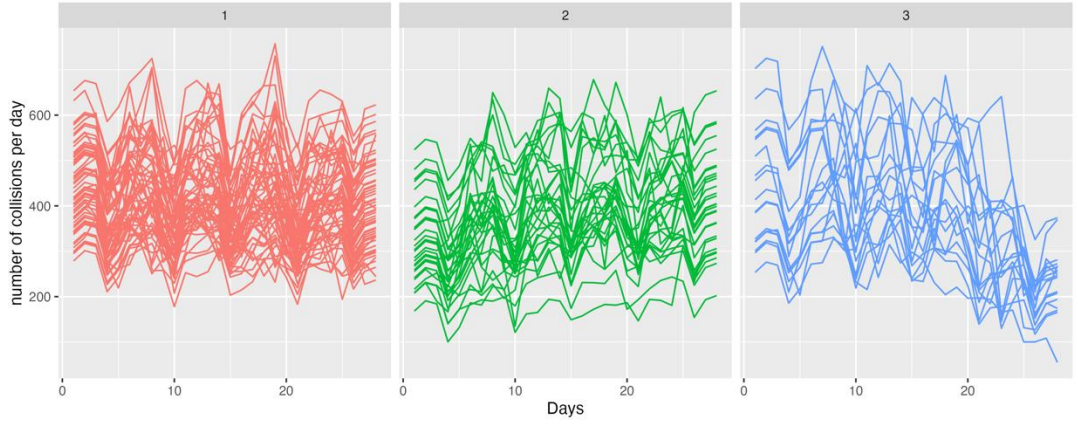
Functional data

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

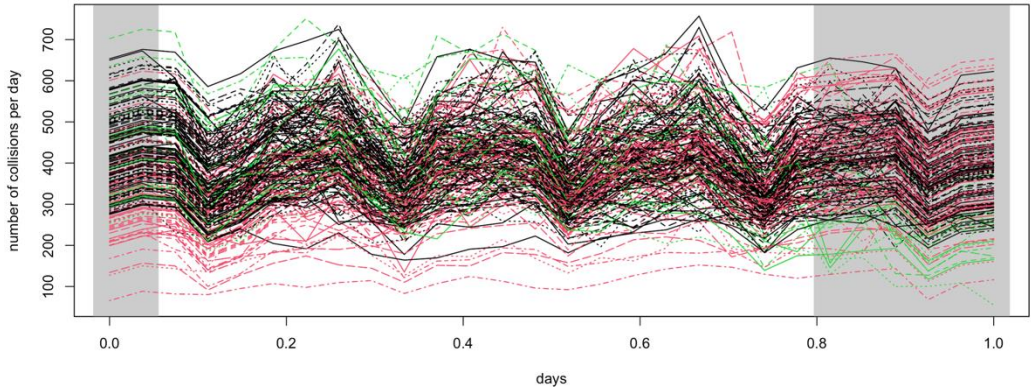
We found **3 clusters**: one containing predominantly the months of **January** and **April**, one containing mostly **December** and the **remaining** months were clustered together.



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:

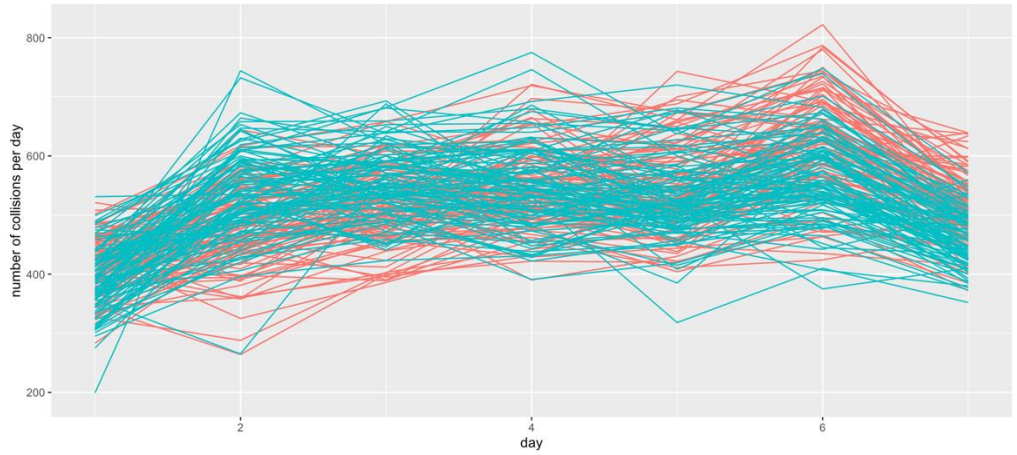
- Regular working weeks;
- Weeks belonging to a holiday period.

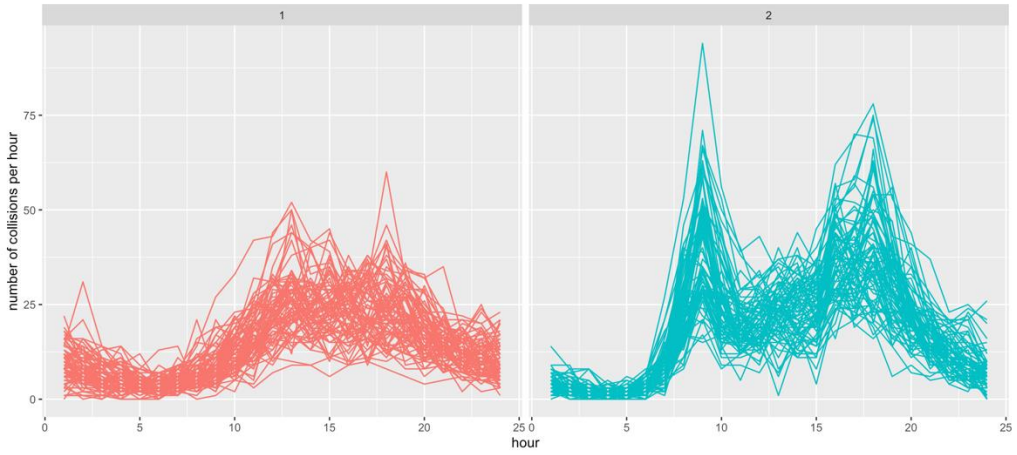
Daily data

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.**





Robust GAM

Robust GAM for the average number of accidents per day

We considered data spanning from 2005 to **2021**. Considering the restrictions on normal vehicular movement during the **COVID** period, we decided to employ a **robust method**.

$$Average\ n^o\ of\ collisions \sim day\ type + wind + rainfall + f_1(day\ of\ the\ year)$$

Rainfall
- Fine 
- Rain 
- Snow 

Day of the year: 1, … , 365
modelled using **natural cubic splines**

Year: 2005, … , 2021

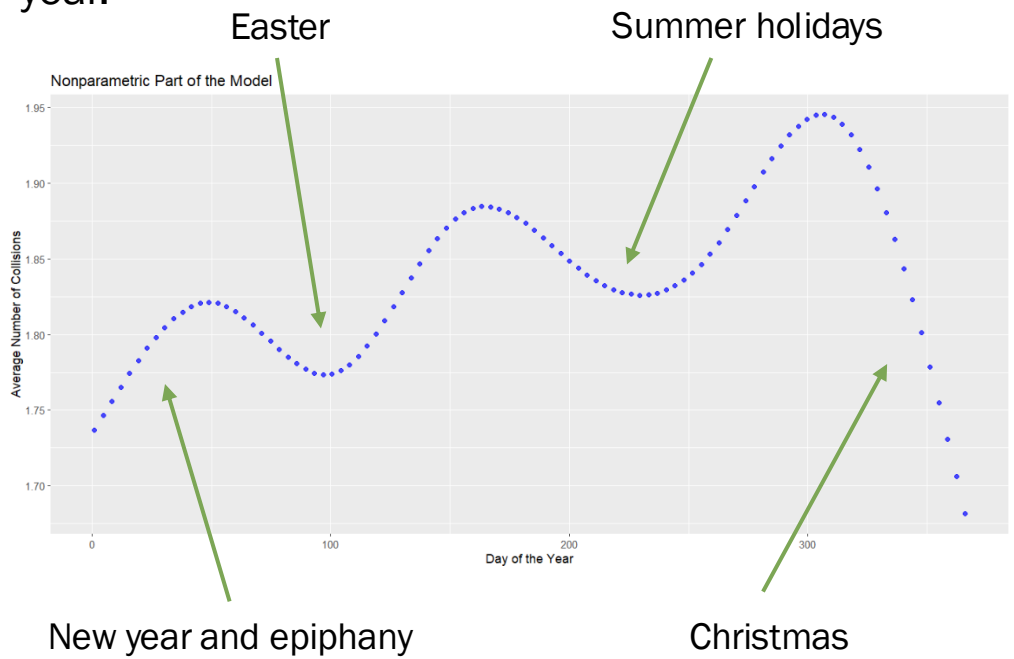
Day type
- Weekday
- Weekend

Wind
- High wind 
- No high wind

Results

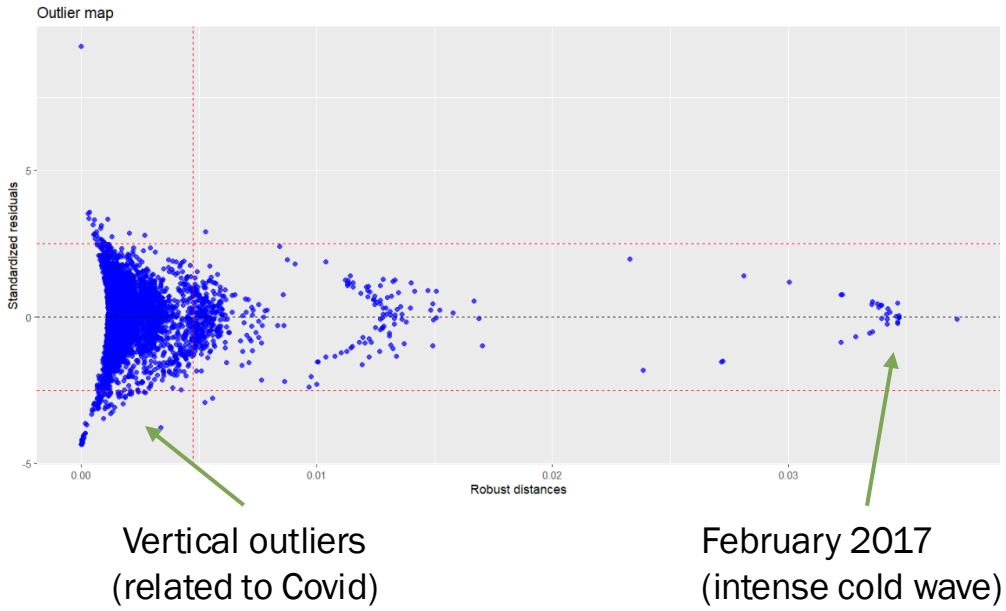
Day of the year

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.



Outlier map

The most recent **vertical outlier** corresponds to the observation on March 24, 2020. The full **lockdown** was officially declared on March 23, 2020, by Prime Minister Boris Johnson.



Conformal prediction

Prediction bands for the daily number of accidents

We decided to produce **conformal prediction** bands following a **split conformal** approach for each district.

The predictions were built from the output of a **GAM** built **independently** for **each district**.

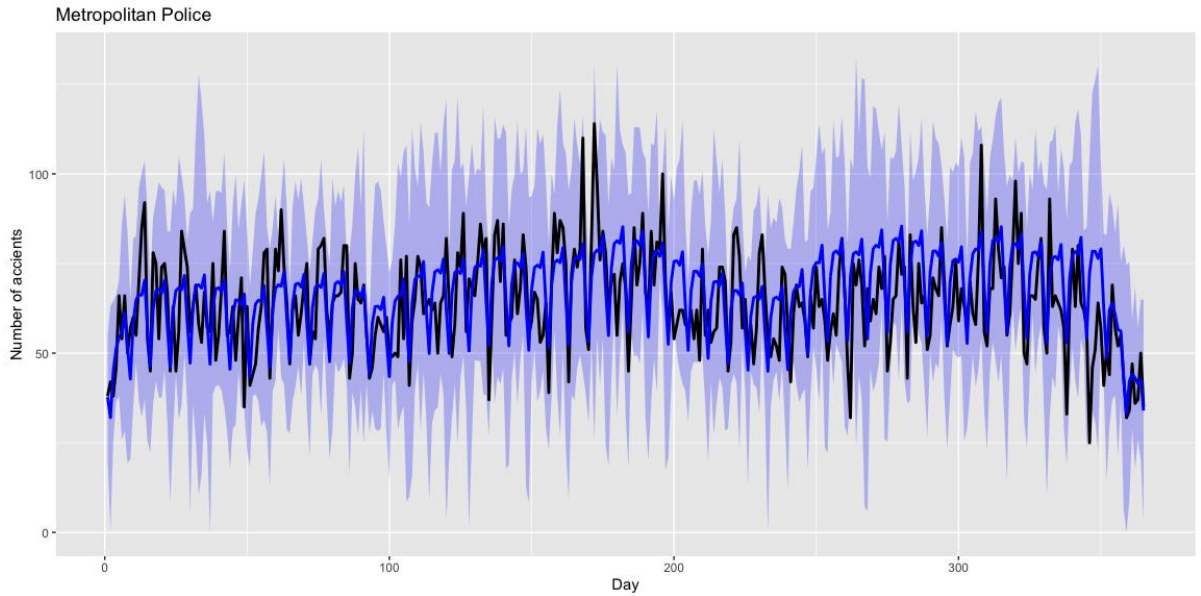
The prediction bands were obtained by:

$$C^s_{n,1-\alpha} := \{y \in \mathcal{Y}(\mathcal{T}) : y(t) \in [g_{\mathcal{I}_1}(t) \pm k^s s_{\mathcal{I}_1}(t)] \forall t \in \mathcal{T}\}$$

where:

$$s_{\mathcal{I}_1}(t) = \frac{\max_{j \in \mathcal{H}_2} \left| y_j(t) - g_{\mathcal{I}_1}(t) \right|}{\int_{\mathcal{T}} \max_{j \in \mathcal{H}_2} \left| y_j(t) - g_{\mathcal{I}_1}(t) \right| dt}$$

$$\mathcal{H}_2 := \{j \in \mathcal{I}_2 : \sup_{t \in \mathcal{T}} \left| y_j(t) - g_{\mathcal{I}_1}(t) \right| \leq k\}$$



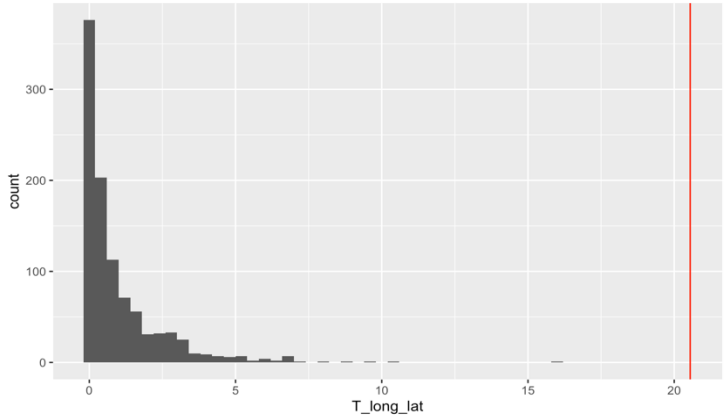
Nonparametric Tests and ANOVA

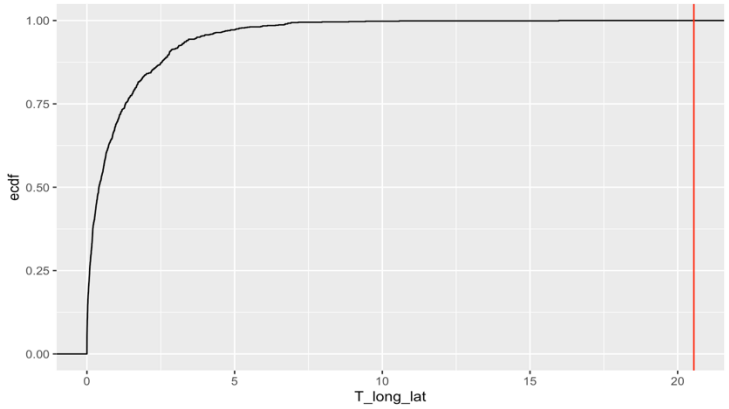
Significance of latitude and longitude on the number of crashes

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
```

- The **interaction** have a **significant** impact on the number of crashes. This suggests that geographic location is an important factor in traffic accidents.
- The p-value of the test: H_0: $interaction_{i,j} = 0$ vs H_1: $interaction_{i,j} \neq 0$ is equal to zero, hence the main effect are significant as well
- **Permutational** ANOVA was used because the data do **not** follow the Gaussian distribution.





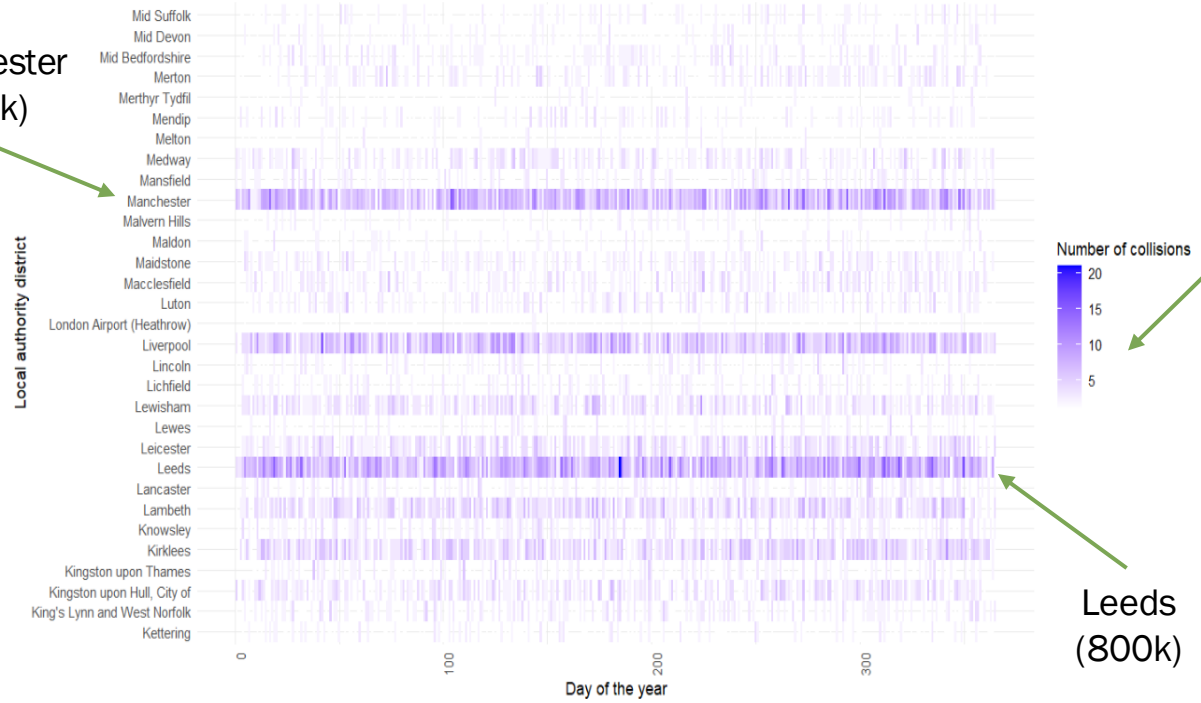
N° of daily collisions per district

GAM model with mixed effects

Year: 2005, … , 2019

$$\log(mean\ n°\ of\ collisions_i) \sim day\ type_i + wind_i + rainfall_i + year_i + f_1(day\ of\ the\ year_i) + b_i$$

Manchester (550k)



Leeds (800k)

Heatmap of the number of collision for 30 districts during the year 2005.

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

Results

Day of the year

We decided not to consider Covid years.

Year: 2005, … , 2019

The nonparametric part of the model is analogous to the one shown in the robust model.

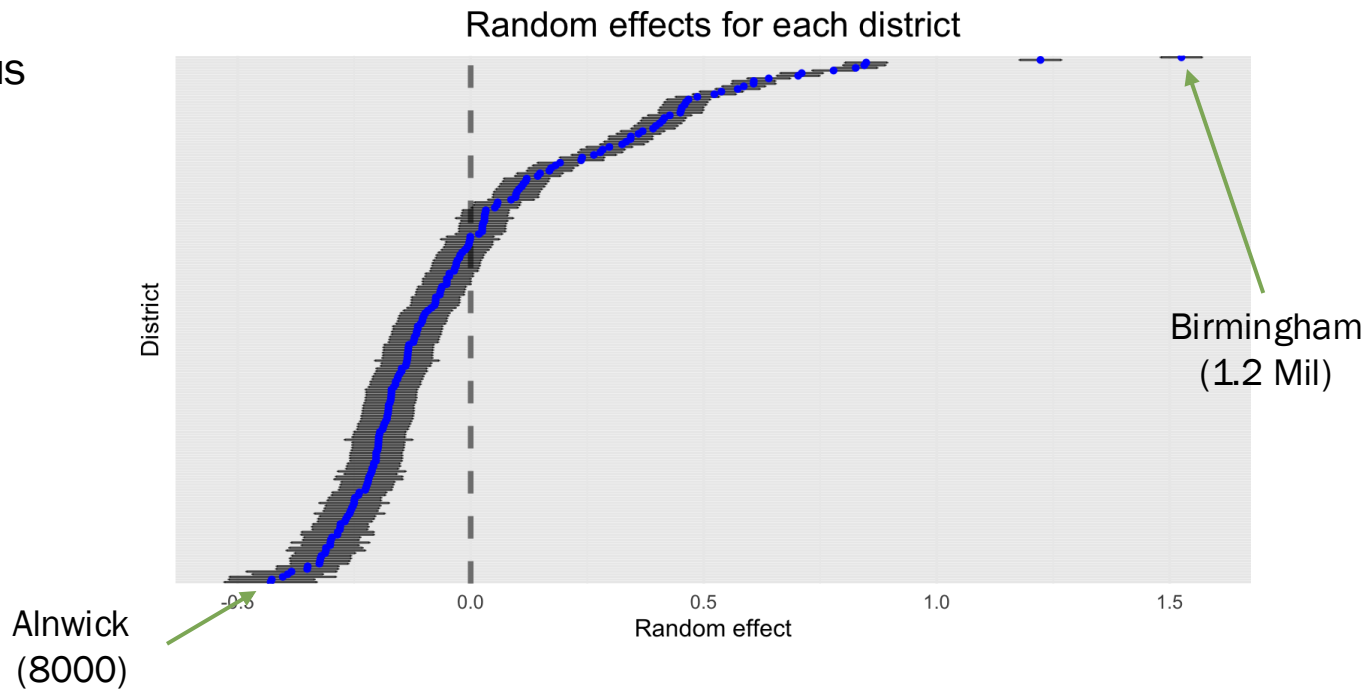
Rain fall

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

Random effects

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



Including spatial information

GAM model using latitude and longitude

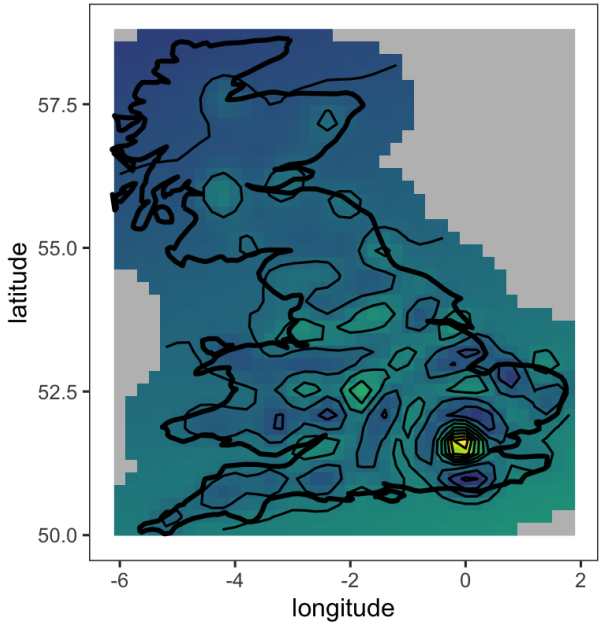
We decided to include in the model the **latitude** and **longitude** components to better model the spatial distribution of the data.

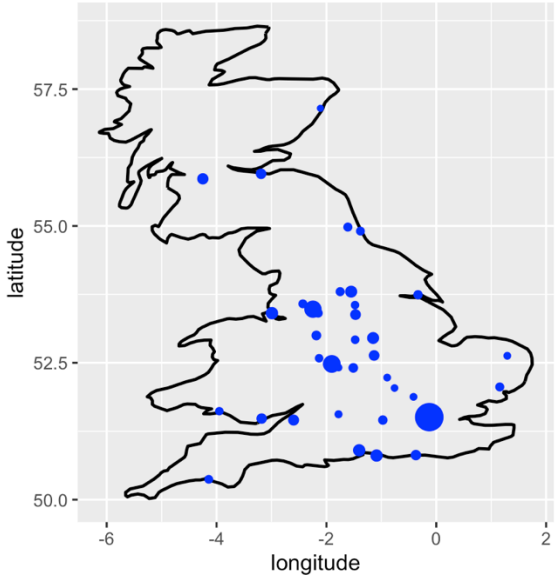
We are capturing the distribution of the **population** on the territory with the added non parametric term.

$$\log(mean\ n°\ of\ collisions) \sim day\ type + wind + rainfall + year + f_1(day\ of\ the\ year) + f_2(longitude, latitude)$$

The spatial part was modelled using a **Gaussian Process** with covariance function:

$$\begin{cases} 1 - 1.5d/\rho + 0.5(d/\rho)^3 \text{ if } d < \rho \\ 0 \text{ otherwise} \end{cases}$$

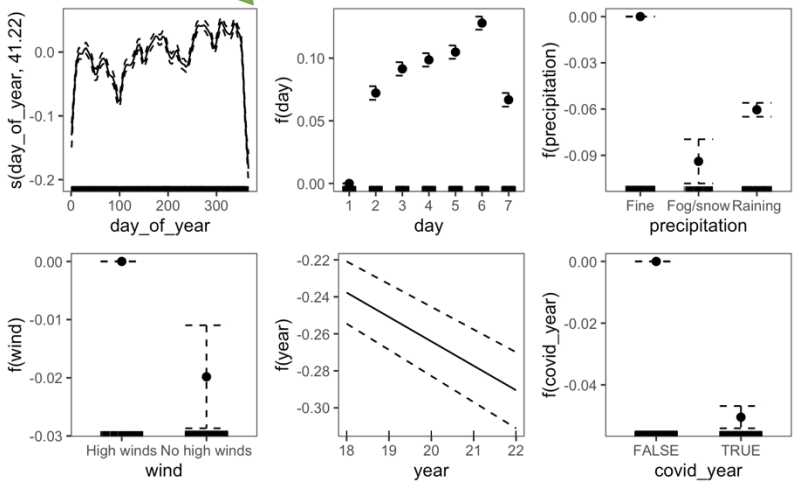


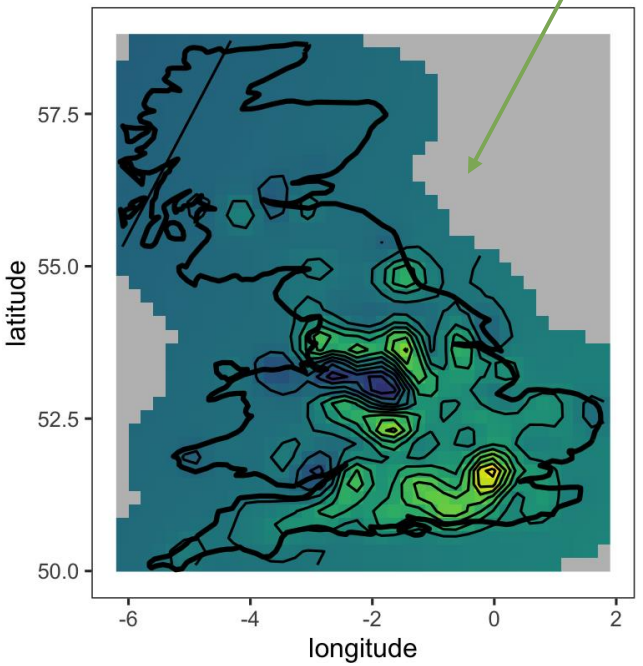


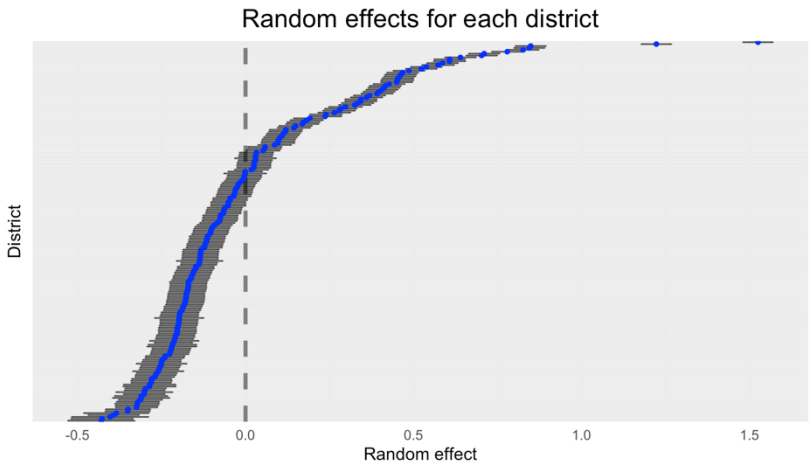
Combining both approaches

The complete GAM model

$$\log(mean\ n°\ of\ collisions_i) \sim day\ type_i + wind_i + rainfall_i + year_i + covid\ year_i + f_1(day\ of\ the\ year_i) + f_2\ (logitude\ _i, latitude_i)\ + b_i$$







Thank you for the attention!

References:

1.UK Department for Transport. Road safety data, from https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

2. J. Diquigiovanni, M. Fontana, and S. Vantini. The importance of being a band: Finite- sample exact distribution-free prediction sets for functional data, 2021.

3. A. Pini and S. Vantini. Interval-wise testing for functional data. Journal of Nonparametric Statistics, 29(2):407–424, 2017.

4. A. Pini, S. Vantini, B. M. Colosimo, and M. Grasso. Domain-Selective Functional Analysis of Variance for Supervised Statistical Profile Monitoring of Signal Data. Journal of the Royal Statistical Society Series C: Applied Statistics, 67(1):55–81, 2017.

5. L. M. Sangalli, P. Secchi, S. Vantini, and V. Vitelli. k-mean alignment for curve clustering. Computational Statistics Data Analysis, 54(5):1219–1233,2010.