POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Project Report

Nonparametric Analysis of UK Road Accidents

Nonparametrtic Statistics (055702)

Authors: Valeria Iapaolo, Oswaldo Morales, Riccardo Morandi, and Abylaikhan Orynbassar

Professors: S. Vantini, F. Ieva, L. M. Sangali, A. Cappozzo

Academic year: 2023-2024

Abstract

This report contains a nonparametric analysis of the traffic accidents and casualties in the United Kingdom in the years from 2005 to 2022. We start by analysing the stakeholders, the first responders, in the form of both emergency services and law enforcement, and we try to provide meaningful information in order to improve their services.
We focus in particular on three distinct aspects:

- investing the factors that influence the severity of a collision;
- modeling the distribution of the number of collisions both in time and space;
- predicting the daily number of accidents in a given district.

The source code related to this project is available at
`https://github.com/oswaldojml/NPS_RoadTrafficCollision`

Contents

1 Problem statement **3**
1.1 Datasets presentation . . . 3
1.2 Stakeholders analysis . . . 3
1.3 Research goals . . . 4
1.4 Exploratory data analysis . . . 4

2 Accident severity **5**
2.1 Permutational ANOVA . . . 5
2.2 GAM model on accident severity . . . 5

3 Number of accidents per day **7**
3.1 Functional Data Analysis . . . 7
3.1.1 Yearly data . . . 7
3.1.2 Monthly data . . . 7
3.1.3 Weekly data . . . 8
3.1.4 Daily data . . . 9
3.2 How to incorporate the spatial information . . . 9
3.2.1 Robust prediction on the average number of accidents including Covid years . . . 9
3.2.2 Conformal prediction on the number of accidents . . . 10
3.2.3 GAM model with random effects for the district . . . 12
3.2.4 Permutational ANOVA for latitude and longitude . . . 15
3.2.5 GAM using latitude and longitude . . . 15
3.2.6 Combining the previous two approaches . . . 16

4 Conclusions **17**

1. Problem statement

Currently, the United Kingdom records a significant number of road accidents every year since fatalities from road accidents rank as the thirtieth leading cause of death. It should be considered that the most common causes involve various types of diseases (e.g., cancer, stroke...). According to the latest data published by the World Health Organization (WHO) in 2020, road accident fatalities in the United Kingdom numbered 2,169 (0.42% of total deaths) and since they are accidental deaths they could be mitigated by taking the right actions.

1.1. Datasets presentation

The datasets were provided by the UK department of transport [2], it included detailed road safety data about the circumstances of personal injury road collisions in Great Britain from 1979, the types of vehicles involved and the consequential casualties. The statistics relate only to personal injury collisions on public roads that are reported to the police, and subsequently recorded, using the STATS19 collision reporting form. This data contains all the non-sensitive fields that can be made public. We decided to limit our analysis to the year 2005-2022 due to a change in the collection methodology of the crashes in the year 2004 and to the size of the dataset that contains more than 2,5 million crashes. The data was available via the `STATS19` R package [3] that also provided useful functions for the preprocessing stage of the raw data. We have three datasets containing information about the accidents, the vehicles involved in each accident and the casualties of each vehicle in each accident, we combined this information in a reduced dataset of the form:

Collisions	
Variable	**Description**
date	date of the crash
day_of_week	day of the week of the crash location
time	time of the crash
longitude	longitude of the crash location
latitude	latitude of the crash location
police_force	police district authority of crash site
accident_severity	severity in 3 categories: Slight,Serious ,Fatal
number_of_vehicles	number of vehicles involved
number_of_casualties	number of casualties in the accident
local_authority_district	district of the location of the collision
road_type	road type: single/dual carriageway, one-way
speed_limit	speed limit: 20,30,40,50,60,70 mph
light_conditions	darkness and presence of lighting on road
weather_conditions	weather conditions:fog, rain, snow etc...
urban_or_rural_area	the area where the accident happened

Knowing the potential number of accidents that can occur in a specific area provides support for the planning of road and healthcare rescue services. It is important, therefore, to identify the time periods during which lethal accidents are most likely to occur and the areas where the incidence is particularly high. It is essential that in these locations and during these time periods, healthcare and police personnel are present in sufficient numbers, organizing shifts based on probable incidents and intensifying checks by the traffic police.

1.2. Stakeholders analysis

To ensure that this report is effective in meeting its intended purpose, it is important to consider the perspectives and needs of the various stakeholders who will be affected by the findings.

- **Emergency medical services:**
  - Resource planning: knowing the potential number of accidents in a specific area helps in planning the allocation of resources for emergency medical services. This includes paramedics and ambulance services.
  - Timely response: identifying high-risk periods and areas allows paramedics to be strategically positioned for a more prompt response, potentially saving lives.
  - Optimizing shifts: it enables the optimization of paramedic shifts based on the probability of incidents, ensuring adequate coverage during critical periods.
- **Law enforcement:**
  - Enhanced safety measures: understanding the likelihood of accidents in specific areas assists the police in implementing targeted safety measures.
  - Strategic presence: identifying high-incidence locations allows the police to strategically position personnel for effective law enforcement and accident prevention.
  - Traffic management: increased awareness of potential accident-prone periods enables the police to intensify traffic checks, ensuring adherence to regulations and reducing the risk of accidents.

1.3. Research goals

Our analysis will primarily focus on addressing two key challenges:

1. **Accident Severity:**
   - Investigate the factors influencing the severity of accidents to identify high-risk scenarios, so authorities can then implement targeted safety measures.
2. **Number of Accidents Per Day:**
   - Examine the variation in the number of accidents to understand temporal patterns;
   - Provide a comprehensive overview of high-incident periods and areas, facilitating effective resource allocation and planning for emergency services;
   - Predict the daily number of accidents in different authority districts.

1.4. Exploratory data analysis

We performed analysis on the different dataset we have, to understand underlying patterns and trends. Due to the categorical nature of the majority of variables, bar plots were employed the most to visualize the distribution of data. Colors were applied to visualize the portion related to an accident severity. In Figure 1 we show some of them.

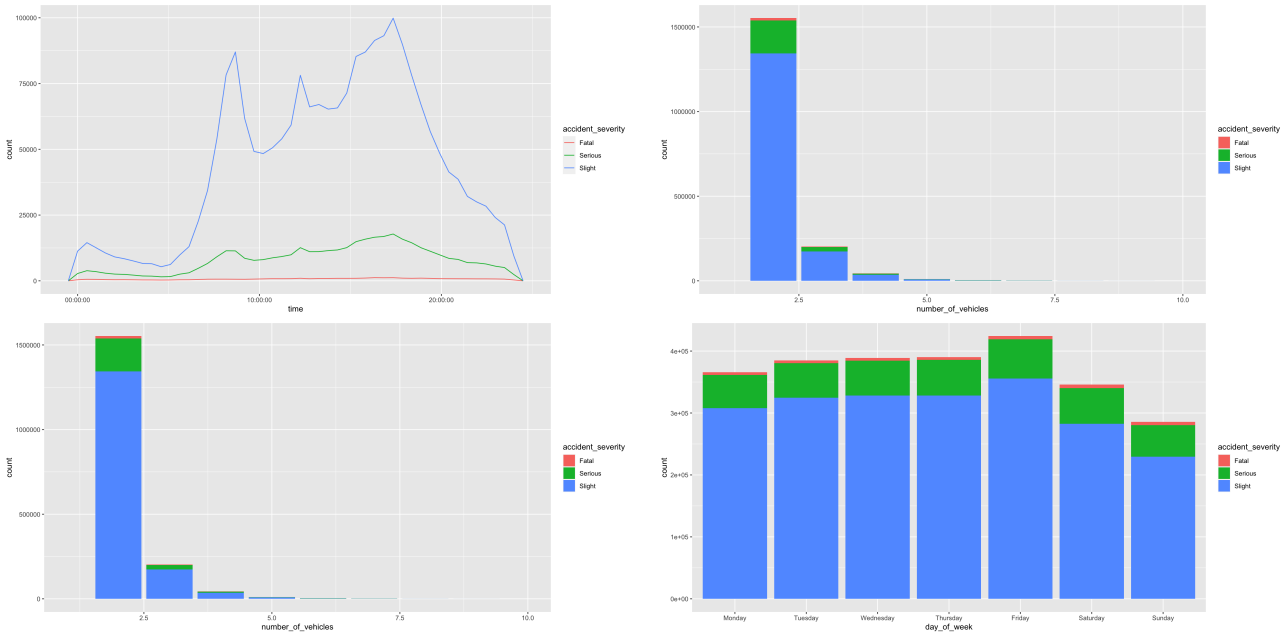

Figure 1: Example EDA on count data and accident severity.

The analysis on accident severity will be carried out in section 2, while the number of accidents will be addressed in section 3 and we will draw some conclusions in section 4.

2. Accident severity

2.1. Permutational ANOVA

Permutational ANOVA was conducted to assess the significance of the number of casualties, the light conditions and the weekend on the casualty severity in the provided dataset. The analysis aimed to determine whether the above mentioned covariates, significantly influence the severity of the injuries in reported accidents. The results of the analysis will provide insights into the relationship between casualty characteristics and the severity of their outcomes, aiding in the understanding of factors contributing to the severity of injuries in road traffic accidents.

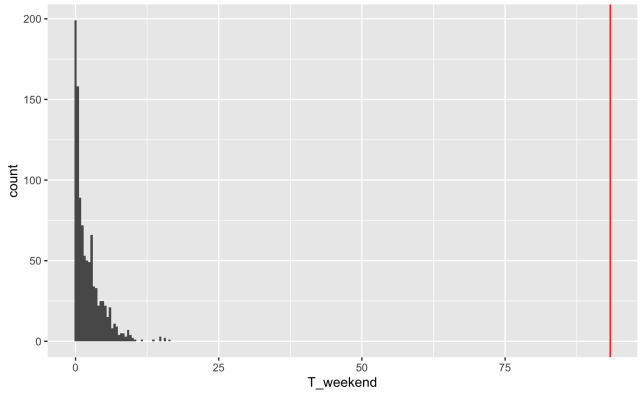

Figure 2: Histogram of the Test Statistic.

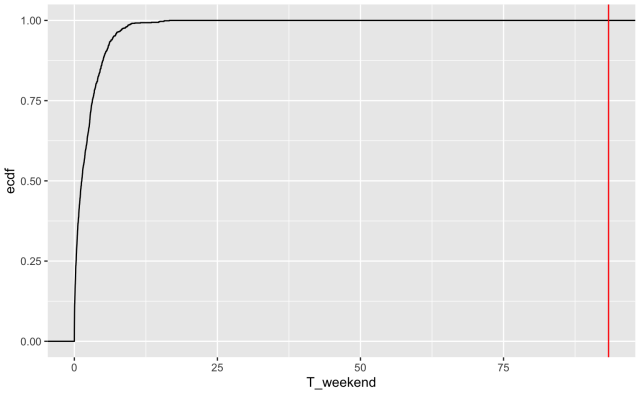

Figure 3: Empirical CDF of the Test Statistic.

The figures 2 and 3 illustrate the histogram of test statistic and ecdf of test statistic for the covariate weekend. The obtained p-value for all the covariates was equal to 0, which suggests that all of them have a significant impact on casualty severity within the dataset. This finding underscores the importance of considering the role of various casualty categories, such as pedestrians, passengers, and drivers or riders, in predicting the severity of injuries resulting from road traffic accidents. The significance of casualty class implies that different groups may experience distinct levels of risk or susceptibility to severe injuries, highlighting the necessity for targeted interventions and safety measures tailored to specific types of road users. Understanding the relationship between casualty class and severity can inform policy decisions, resource allocation, and road safety initiatives aimed at mitigating the risk of serious injuries and improving overall road safety outcomes.

2.2. GAM model on accident severity

In order to analyse the accident severity, we employed two GAM models. For both models we considered the same covariates which are: the number of casualties, time, weekend, number of vehicles, speed limit, light conditions. A portion of the dataset was reserved for model testing purposes.

Both models used the same semiparametric model for the covariates, but with a different way of modelling the response variable. The nonparametric functions are defined as following:

- f_1 is a periodic cubic regression spline
- f_2 is a cubic regression spline
- f_3 is a cubic regression spline

The first one is a GAM model with ordered categorical family [11] which is able to model the three types of categories that accident severity can assume (i.e.: slight, serious and fatal).
The model was specified as follows:

$$\begin{aligned}\text{accident severity}_i \sim \text{number of casualties}_i + \text{weekend}_i + \text{light conditions}_i \\ + f_1(\text{time}_i) + f_2(\text{number of vehicles}_i) + f_3(\text{speed limit}_i)\end{aligned}$$

The imbalanced nature of the data resulted in a prediction table in Figure 4 that exhibited a bias towards the class "slight severity".
In order to address this issue and enhance the model's predictive performance, we opted to merge labels serious severity and fatal severity into a single category and then we fitted a second model, which is a GAM with a logit link function where the response variable is "slight severity" or "serious or fatal severity".

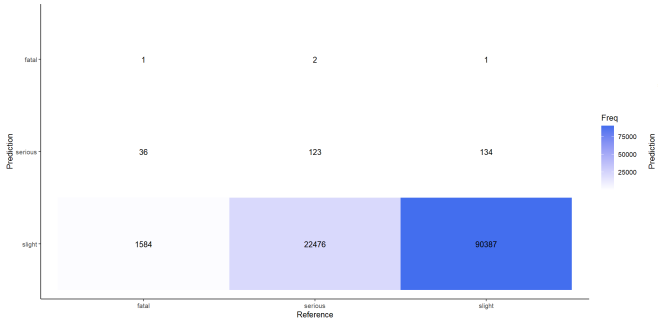

Figure 4: Prediction table of the model with ordered categorical family.

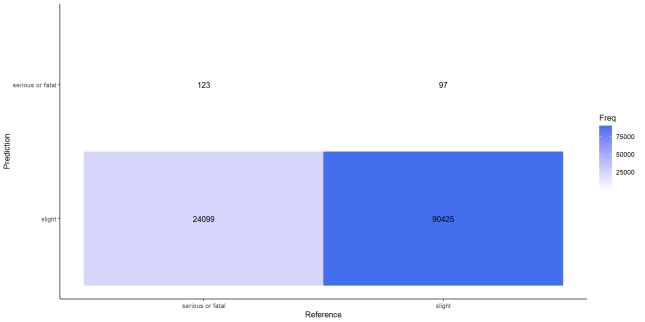

Figure 5: Prediction table of the model with binary response.

The model with the binary response variable had the following form:

$$\log(\text{accident severity}_i) \sim \text{number of casualties}_i + \text{weekend}_i + \text{light conditions}_i + f_1(\text{time}_i) + f_2(\text{number of vehicles}_i) + f_3(\text{speed limit}_i)$$

The coefficients of the parametric part of the model are reported in Table 1, the fitted nonparametric components are depicted in Figure 6 and the predition's confusion matrix is visible in Figure 5.

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

Table 1: Coefficients of the parametric part of the model with binary response.

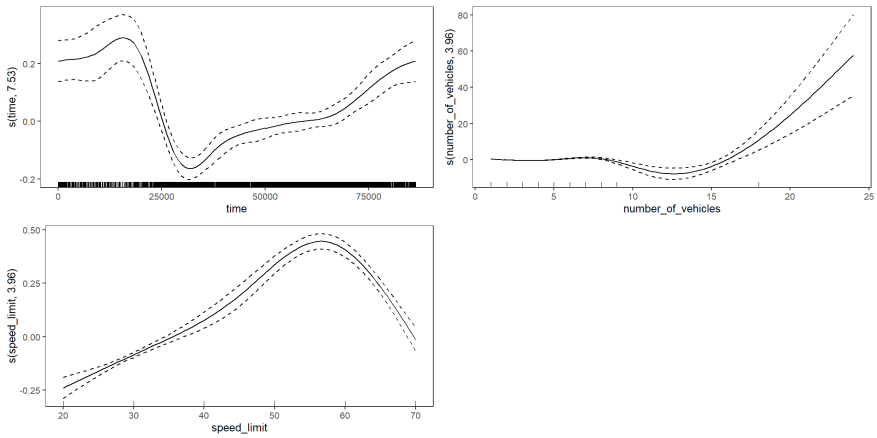

Figure 6: splines of the model with binary response.

All parametric coefficients in our Generalized Additive Model (GAM) for accident severities with a binary response demonstrate a positive influence on the severity of collisions. Notably, the dummy variable representing 'light conditions: Daylight' did not exhibit statistical significance in the parametric test conducted by the model. While all other parameters demonstrated significance, it is prudent to consider conducting nonparametric tests to further validate the significance of these variables.
Examining the splines, our analysis reveals that from night until early morning, there is a positive impact on collision severity. Additionally, the number of vehicles and speed limit both contribute to an increased severity; however, a noteworthy observation is that beyond a speed limit of 60 mph, the impact on severity begins to decrease. The reasons behind this diminishing effect warrant further investigation.

3. Number of accidents per day

3.1. Functional Data Analysis

We first started by representing the total number of accidents in the whole UK as a functional data, considering 4 different time frames: years, months, weeks and days. In the first three cases we considered the total umber of daily cashes, while in the daily case we group the accidents by hour.

3.1.1. Yearly data

When considering the number of crashes in each year we found that the number of crashes per year decreased over time year after year, even though the total number of cars sold in the UK increased in each of the years considered in this project. The decrease appears to be gradual in the years and we do not see a structure in the data that cold lead to a successful clustering.
Furthermore using a functional bagplot [5] it is clearly visible the outlying year 2020, which is well below the bag due to the lockdowns during the pandemic.

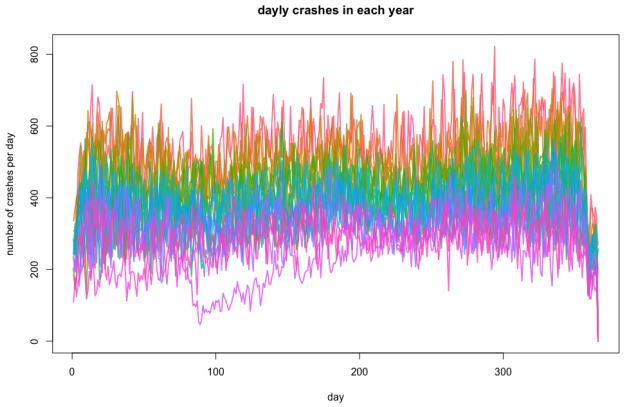

Figure 7: daily crashes in each year.

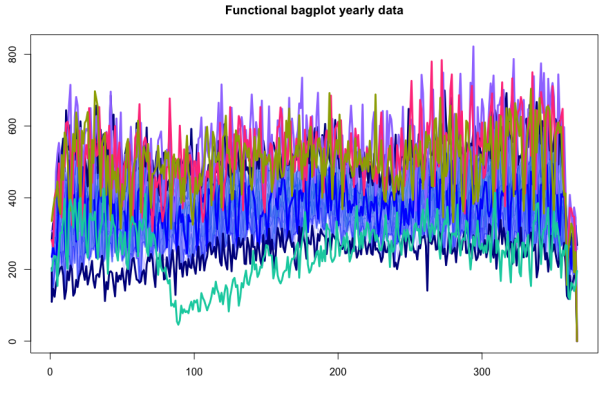

Figure 8: Functional bagplot using F=1.5.

3.1.2. Monthly data

In the case of the monthly data we decided to align the data using a shifting warping function to properly capture the weekly pattern in the data, after the alignment a clear pattern emerged in the data.
We performed functional clustering using the functional k-means algorithm [9], implemented in `fdacluster` [10], with no alignment, since we aligned the curves beforehand, using Pearson's distance, the mean curve as centroids and setting k = 3. The resulting clusters are displayed in figure 9:

1. containing the majority of the months.
2. containing mostly the months of January and April;
3. containing mostly the month of December;

The statistical significance of the resulting clusters was tested using global permutational tests using as test statistics the L^2 distance between the curves.

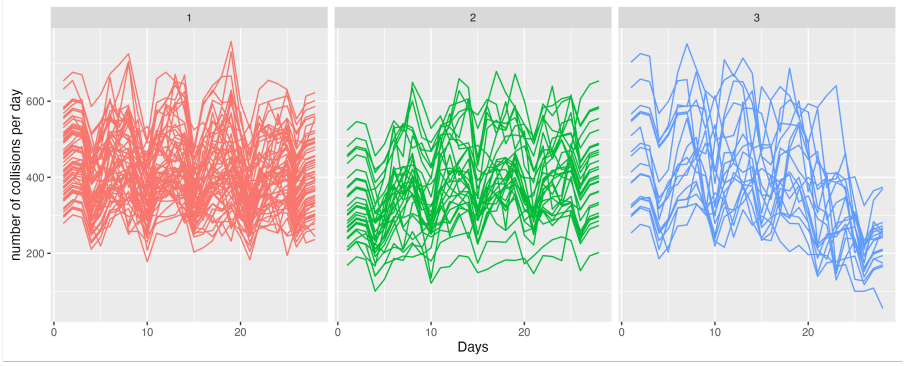

Figure 9: clusters obtained by functional k-means.

The difference between the clusters was concentrated at the beginning an at the end of the month, therefore we decided to perform Interval Wise Testing [6], an inferential procedure for functional data able to select the intervals of the domain imputable of rejecting a functional null hypothesis. Since we had three groups we performed interval wise Functional Analysis of Variance [8], this method was already implemented in the `fdatest` [7] R package.
The functional ANOVA model that we considered is:

$$y_{i,j}(t) = \mu(t) + \alpha_i(t) + \varepsilon_{i,j}(t)$$

where $\mu(t)$ is the functional grand mean, $\alpha_i(t)$ is the functional main effect ad $\varepsilon_{i,j}(t)$ are assumed to be independent and identically distributed zero-mean random functions. We want to test

$$\boldsymbol{H_0} : \alpha_i(t) = 0 \quad \forall i \in \{1,2,3\} \qquad \text{v.s.} \qquad \boldsymbol{H_1} = (\boldsymbol{H_0})^C$$

The results of the functional Analysis Of Variance are represented in figure 10 and it is clear that the difference is significant only at the extremities of the time domain as indicated by the dark shading of the figure representing the adjusted p-value. This behaviour can be easily explained by the presence of holidays: the decrease at the beginning of January is due to new years and epiphany, the one at the beginning of April due to Easter, and the one at the end of December toe to the Christmas holidays.

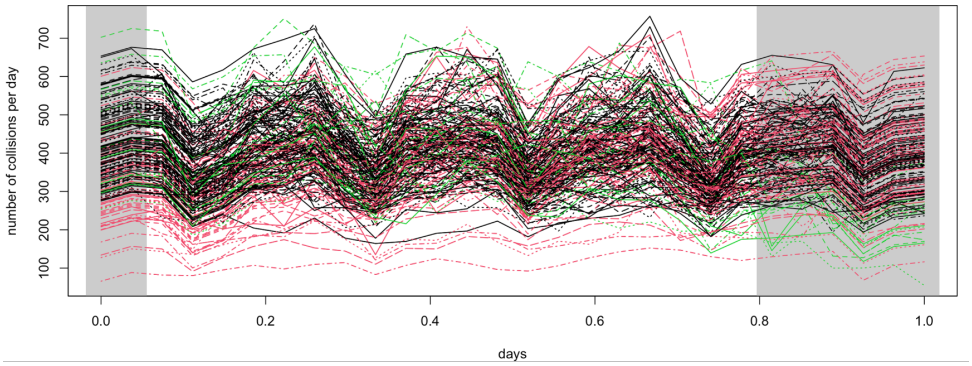

Figure 10: one way functional analysis of variance on the clusters.

3.1.3. Weekly data

We performed a similar procedure to the one outlined in section 3.1.2 for the weekly data, in this case there was no need for an alignment. We obtained two distinct clusters one corresponding to working weeks and one to holiday week as visible in figure 11.
As before we performed an global permutation test that validated the statistical significance of the clusters. To further investigate the difference between the clusters we performed the following Interval Wise test [6] on the two clusters, with $\boldsymbol{Y_i}$ restricted to I:

$$\boldsymbol{H_0} : \boldsymbol{Y_1^I} = \boldsymbol{Y_2^I} \qquad \text{v.s.} \qquad \boldsymbol{H_1} : \boldsymbol{Y_1^I} \neq \boldsymbol{Y_2^I}$$

For any t in T we define the IWT-adjusted p-value function, show in Figure 12, as:

$$\hat{p}(t) = \sup_{t \in I} p'$$

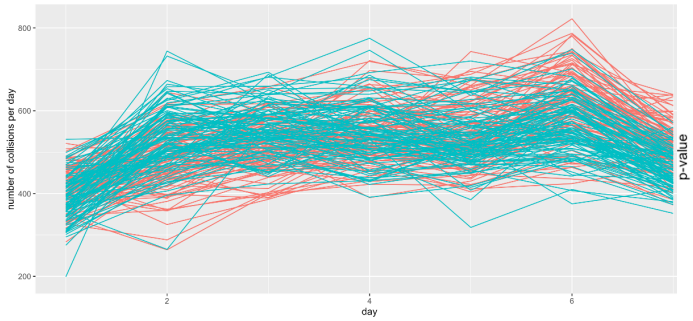

Figure 11: clusters obtained by functional k-means.

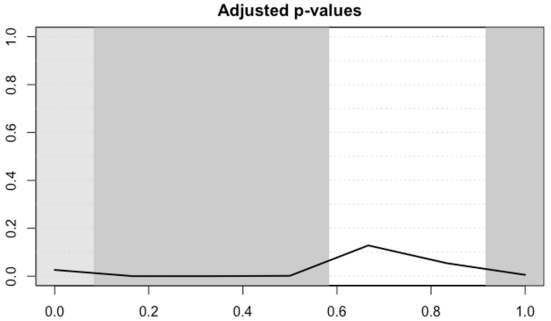

Figure 12: adjusted p-value function.

3.1.4. Daily data

We performed a similar procedure to the one outlined in section 3.1.3 for the daily data,this time using the total number of accidents per hour. We obtained two distinct clusters one corresponding to working days, where two peaks corresponding to rush hour traffic are visible, and one to holidays, where the accidents are overall much fewer, as shown in figure 11.
A global permutation test that validated the statistical significance of the clusters, while an Interval Wise Test analogous to the one in section 3.1.3 revealed that the difference is significant during the whole time period.

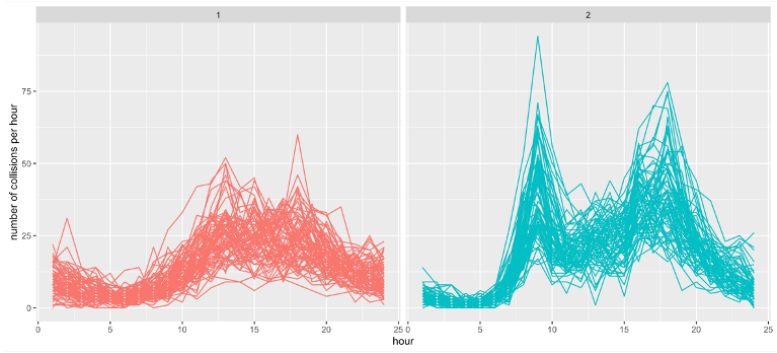

Figure 13: one way functional analysis of variance on the clusters.

3.2. How to incorporate the spatial information

We have explored various strategies to incorporate the spatial information into our analysis. The first approach, as outlined in Section 3.2.1, is the simplest: we calculate the average number of collisions per day by averaging across districts. In Section 3.2.2, we employ conformal prediction by considering each district individually. Subsequently, in Section 3.2.3, we introduce a random intercept for each district into our model. Following an assessment of the significance of the latitude and longitude variables in Section 3.2.4, we utilize them in Section 3.2.5. Finally, in Section 3.2.6, we integrate the approaches from Sections 3.2.3 and 3.2.5.

3.2.1. Robust prediction on the average number of accidents including Covid years

In this section, our emphasis is on determining the average daily number of accidents, utilizing robust statistics and averaging across districts. This was accomplished by taking into account data spanning from 2005 to 2021. Considering the restrictions on normal vehicular movement during the COVID period, we decided to employ a robust method. We aim to construct a semiparametric model that incorporates factors such as the day type ("weekend"/"weekday"), the year and the weather conditions in the parametric component. Additionally, we intend to include the day of the year in the non-parametric segment, employing natural cubic splines with 7 degrees of freedom.

$$average\, n^\circ\, of\, collisions_i \sim \text{day type}_i + \text{wind}_i + \text{rainfall}_i + \text{year}_i + f_1(\text{day of the year}_i)$$

The model was fitted using the R package `robustbase` [4] obtaining $R^2_{\text{adj}} = 0.502$ and the coefficients for the parametric part shown in Table 2.

Coefficients of the parametric part	
Intercept	44.632
Dummy for the category "Weekend"	-0.183
Dummy for the category "No high winds"	-0.002
Dummy for the category "Fog"	-0.030
Dummy for the category "Raining"	0.007
Dummy for the category "Snowing"	-0.032
Year	-0.021

Table 2: Coefficients of the parametric part.

Now, we focus on the non parametric part. From Figure 14 we clearly see a noticeable decrease in the number of collisions during Easter, summer holidays, and in the period from Christmas holidays to Epiphany.

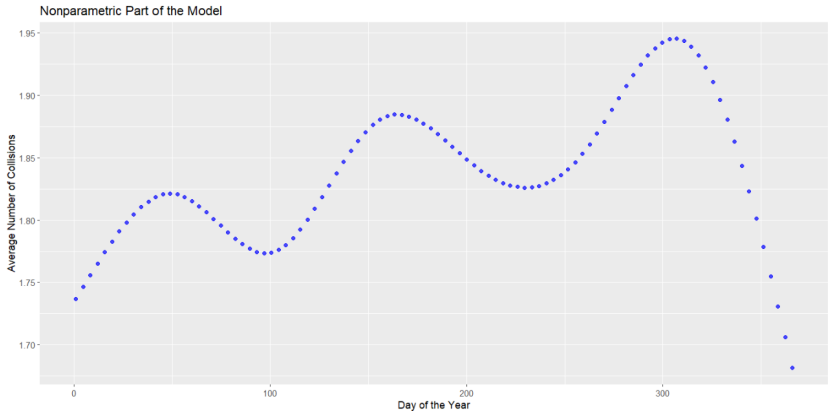

Figure 14: Plot of the non linear function f_1 capturing the effect of the day of the year on the average number of collisions per day.

From the outlier map depicted in Figure 15, we discern the presence of vertical outliers, good leverage points, and bad leverage points. Upon identifying the indices of observations categorized as vertical outliers, it becomes apparent that these are associated with the COVID period. Specifically, the initial 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. On this day, the UK government implemented a series of restrictions, encompassing the closure of numerous non-essential businesses, constraints on people's movements and the prohibition of social gatherings. It is noteworthy that among the observations classified as good leverage points, there are data pertaining to February 2017, during which the UK underwent an unusually prolonged and intense cold wave, accompanied by substantial snowfall, particularly in northern and eastern regions.

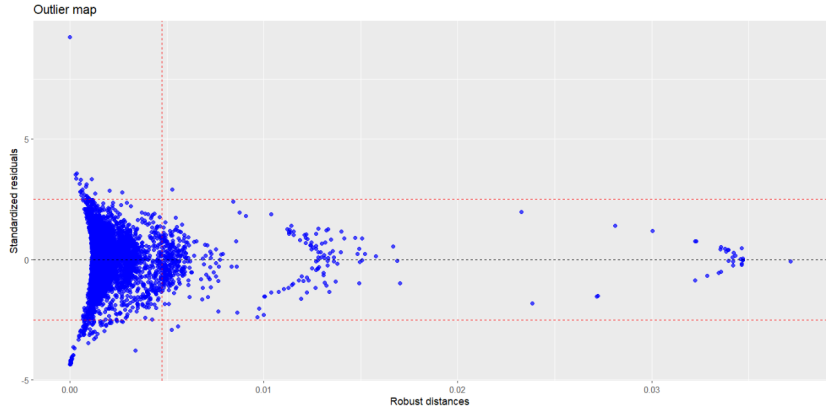

Figure 15: Outlier map for robust model.

Now, let's delve into why it is advantageous to employ a robust method in this scenario. Fitting the same model without robust methods yields residuals that do not follow a normal distribution. Conversely, when utilizing the robust model after identifying outliers through the utilization of the outlier map, the normality of the residuals must be satisfied by the majority of observations, excluding the outliers. This precisely characterizes the situation in this case as shown in Figure 16.

3.2.2. Conformal prediction on the number of accidents

At this point we decided to predict the number of accidents per day, in particular we decided to build a separate Generative additive model for each of the district and use its predictions to build prediction bands. This can be obtained with Conformal Prediction, a non parametric approach that generates prediction sets under the only assumption of exchangeable regression pairs.

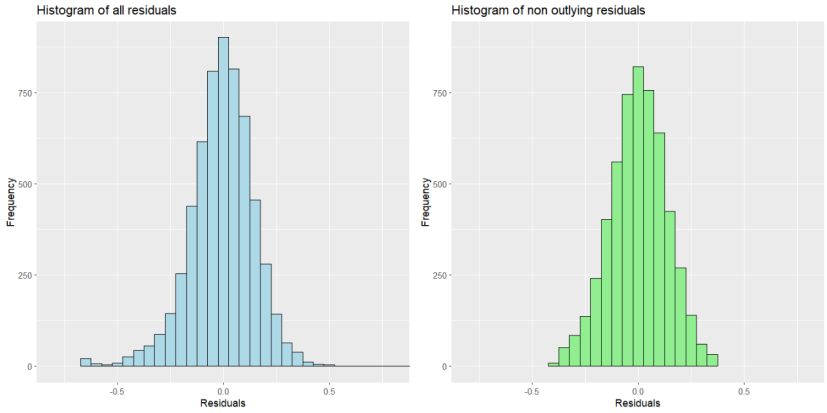

Figure 16: Histogram of all the residuals (on the left) and histogram of non outlying residuals (on the right).

Given a set of i.i.d. random functions $\mathbf{Y_1}, \ldots, \mathbf{Y_n} \sim P$ and an independent random function $\mathbf{Y_{n+1}} \sim P$, a valid prediction set $C_{n,1-\alpha} := C_{n,1-\alpha}(\mathbf{Y_1}, \ldots, \mathbf{Y_n})$ for $\mathbf{Y_{n+1}}$ is such that:

$$\mathbb{P}(\mathbf{Y_{n+1}} \in C_{n,1-\alpha}) \geq 1 - \alpha$$

for any significance level $\alpha \in (0, 1)$, with $\mathbb{P}$ representing the probability corresponding to the product measure induced by P.
In particular, we applied Split Conformal prediction: the data $y_1, \ldots, y_n$ is randomly divided into two sets $\mathcal{I}_1, \mathcal{I}_2$, the training set is defined as $\{y_h : h \in \mathcal{I}_1\}$ while the calibration set is $\{y_h : h \in \mathcal{I}_2\}$.
Following the approach presented in [1], we adopt the same nonconformity measure and modulation function. We want prediction bands that adapt their width according to the local variability of functional data, so the nonconformity measure used is:

$$A(\{y_n : h \in \mathcal{I}_1\}, y) = \sup_{t \in \mathcal{T}} \left| \frac{y_(t) - g_{\mathcal{I}_1}(t)}{s_{\mathcal{I}_1}(t)} \right|$$

and the nonconformity scores are:

$$R_j^s = \sup_{t \in \mathcal{T}} \left| \frac{y_j(t) - g_{\mathcal{I}_1}(t)}{s_{\mathcal{I}_1}(t)} \right| \qquad R_{n+1}^s = \sup_{t \in \mathcal{T}} \left| \frac{y_(t) - g_{\mathcal{I}_1}(t)}{s_{\mathcal{I}_1}(t)} \right|$$

with $j \in \mathcal{I}_2$ and $s_{\mathcal{I}_1} = s(\{h : h \in \mathcal{I}_1\})$.
The split conformal prediction band induced by the nonconformity measure is:

$$C_{n+1,\alpha}^s = \{y \in \mathcal{Y}(\mathcal{T}) : y(t) \in [g_{\mathcal{I}_1}(t) \pm k_{\mathcal{I}_1}^s(t)] \forall t \in \mathcal{T}\}$$

with k^s the $\lceil (u+1)(1-\alpha) \rceil$ th smallest value in $\{R_s^h : h \in \mathcal{I}_2\}$.
In our application $\alpha = 0.1$, $g_{\mathcal{I}_1}(t)$ is defined as the prediction obtained by the GAM, and the modulation function is:

$$s_{\mathcal{I}_1}(t) = \frac{\max_{j \in \mathcal{H}_2} |y_j(t) - g_{\mathcal{I}_1}(t)|}{\int_{\mathcal{T}} \max_{j \in \mathcal{H}_2} |y_j(t) - g_{\mathcal{I}_1}(t)| \, dt} \qquad \mathcal{H}_2 := \{j \in L_2 : \sup |y_j(t) - g_{\mathcal{I}_1}(t)| < k\}.$$

The predictions were built using a GAM with Poisson family and model:

$$\log(\text{mean n° of collisions}_i) \sim \text{day the week}_i + \text{year}_i + f_1(\text{day of the year}_i)$$

where f_1 is a non parametric term based on cubic regression splines.
An example of the obtained prediction bands for the Metropolitan Police department i the year 2022 can be seen in Figure 17, where in black you have the real data, in blue the point prediction from the GAM and the shaded light blue area represents the prediction band.

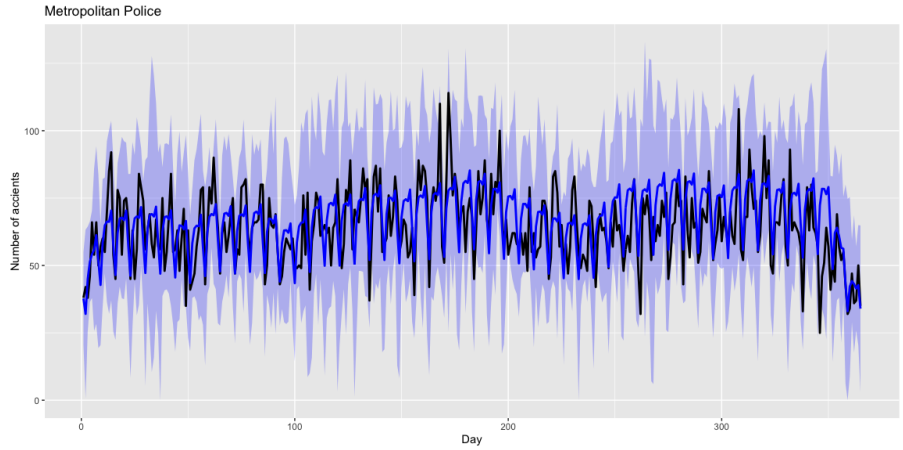

Figure 17: Prediction of the number of crashes for the Metropolitan Police department.

3.2.3. GAM model with random effects for the district

The primary focus of our analysis is to study and potentially predict the number of accidents that may occur in a specific area. We aim to understand which variables are significant in this task.
To achieve this goal, we utilized the dataset of the collisions containing information on every accident that occurred in the UK between 2005 and 2019. It's important to note that, for this part of the analysis, data related to the COVID-19 years was excluded. The data were grouped by date and local authority district to obtain the number of accidents per day and per district. During the grouping process, certain choices were necessary. For instance, to determine the meteorological conditions on a specific day in a particular area, we considered the most frequent class. These adjustments resulted in a new dataset structured as follows:

Dataset for the number of collisions per day and per district	
Variable	**Description**
date	date of the accident
local_authority_district	416 categories: district in which the accident occurred
number_of_collisions	number of accidents in a specific district and in a specific date
most_common_wind	2 categories: "No high wind", "High winds"
most_common_rainfall	4 categories: "Fine", "Raining", "Snowing", "Fog"
year	categorical: year of the collision $\in [2005, \ldots, 2019]$
day_of_the_year	day of the year in which the collision occurred $\in [1, \ldots, 365]$
day_type	2 categories: "Weekday", "Weekend"

Table 3: Dataset for the number of collisions per day and per district.

Clearly there is a correlation between the number of collisions in a district and the number of inhabitants in said district as visible in Figure 18: it is evident that Leeds (about 800.000 inhabitants), Liverpool (about 450.000 inhabitants) and Manchester (about 500.000 inhabitants) are the cities in which more accidents happen. Looking at the heatmap in which all the districts are present, this consideration can be extended to Birmingham, Bradford, Glasgow City, Sheffield, Westminster and Wirral. We would like to note that London does not appear as one would expect since it is divided into multiple districts.
With the aim of understanding which variables are significant for the number of accidents, we fitted a Generalized Additive Model (GAM). The response variable was the number of collisions per day and per district and the covariates included meteorological conditions, year, and day type (as a parametric effect). Additionally, we incorporated the day of the year as a non-parametric effect using cubic splines and treated local authority district as a random effect.

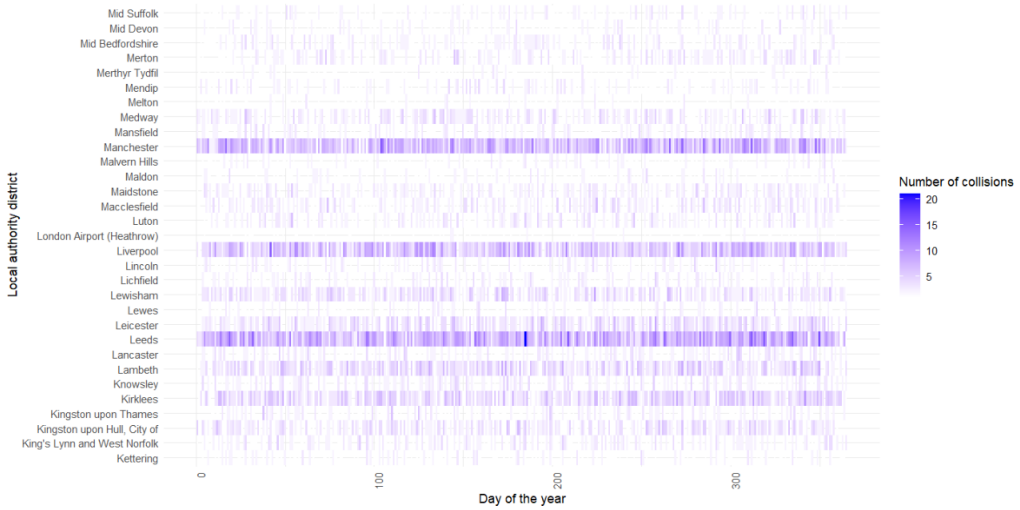

Figure 18: Heatmap of the number of collision for 30 districts during year 2005.

Given that the response variable is a count variable, we opted for the Poisson family for the model. For $i = 1, \ldots, n$, the model is:

$$\log(\text{mean n}^\circ \text{ of collisions}_i) \sim \text{day type}_i + \text{wind}_i + \text{rainfall}_i + \text{year}_i + f_1(\text{day of the year}_i) + b_i$$

where n is the total n° of observations (about 580.000), f_1 is a non linear function and b_i is a random intercept for the local authority district used to capture the difference in the population across the country.
After fitting the model, we obtain coefficients that, at first glance, have a counterintuitive sign. In fact, the signs of the dummy variables for rain and snow are negative, which means that accidents decrease when it rains and even more when it snows. This is reasonable if you consider that adverse weather conditions often lead to reduced outdoor activities, resulting in less use of cars or other vehicles. As an example, in Birmingham the 1st of Feb 2019 the predicted mean number of crashes are:

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

Table 4: Predictions for the mean n° of collisions in Birmingham on the 1st of Feb 2019.

As for the wind intensity, we consistently find that when the wind is stronger, there are more accidents. Conversely, the coefficient for the year is negative: this indicates that over the years, the number of accidents has decreased (this was already observed in the exploratory data analysis phase). In Table 5 there are the coefficients of the parametric part. Regarding the non-parametric component, we obtain similar results to those described in Section 3.2.1.

Coefficients of the parametric part	
Intercept	33.585
Dummy for the category "Weekend"	-0.125
Dummy for the category "No high winds"	0.038
Dummy for the category "Fog"	-0.175
Dummy for the category "Raining"	-0.053
Dummy for the category "Snowing"	-0.089
Year	-0.016

Table 5: Coefficients of the parametric part.

Now let's examine the impact of districts:

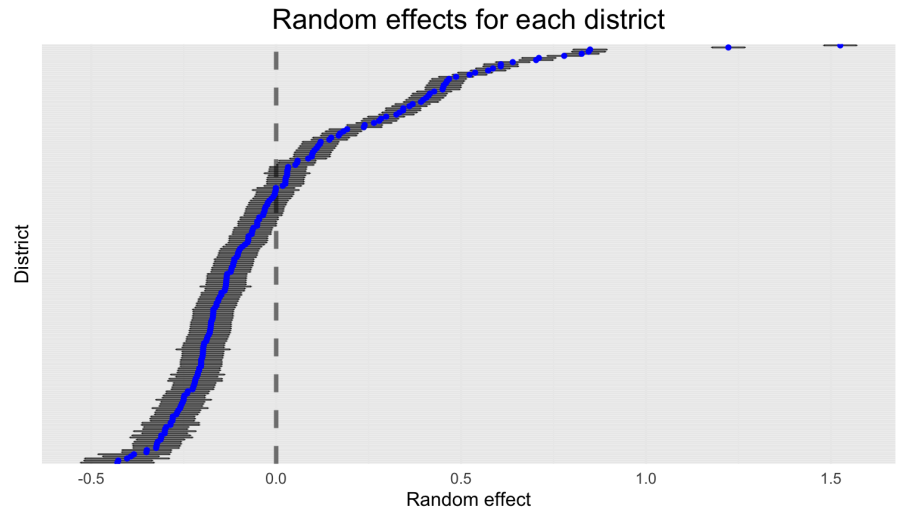

Figure 19: Dotplot of the random effects of the local authority districts.

Of course, the random intercepts for more populated districts have high values, contributing to an increase in the average number of incidents. Similarly, smaller districts are associated with a random intercept having a negative value, as shown in Table 5. The smallest random intercept corresponds to the Alnwick district, with approximately 8,000 inhabitants, on the other hand, the largest corresponds to Birmingham, which has a population of 1.2 million inhabitants.
Unfortunately, the random effects do not follow a normal distribution, leading to a violation of the model assumptions. This circumstance may result in imprecise and unreliable estimates, potentially leading to incorrect conclusions about the relationship between variables. This violation is highlighted in Figure 20 that represents a DD-plot for the random effects and data normally distributed, with the mean and variance aligning with the empirical mean and variance of the random effects. Deviation from a straight line in the plot indicates discrepancies between the two distributions. The same conclusion can be drawn by visually inspecting the histogram of the random effects, the Q-Q plot and applying the Shapiro test.

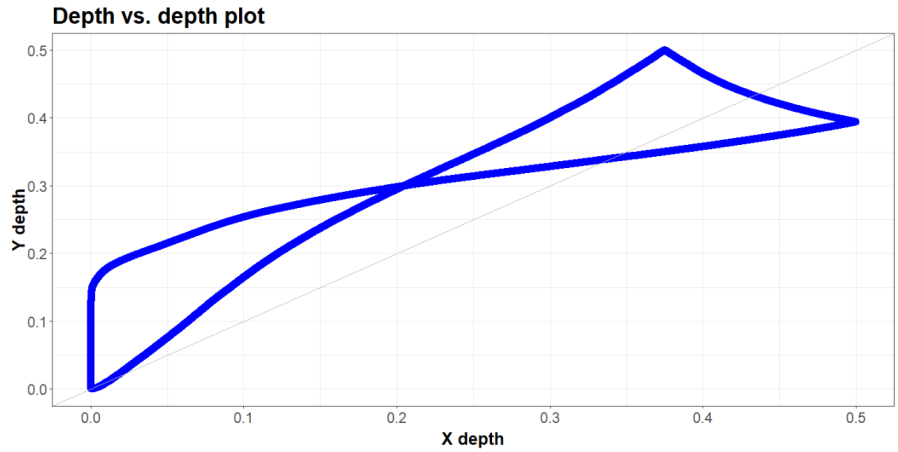

Figure 20: DD-plot for the random effects and data normally distributed, with the mean and variance aligning with the empirical mean and variance of the random effects.

The deviance explained by the model is 40.6%, which leads us to consider the need for a proper incorporation of the spatial components. The analysis conducted so far does not fully leverage the spatial information available, namely the longitude and latitude of the incident location. In this initial attempt, we grouped data by districts, resulting in areal data.

3.2.4. Permutational ANOVA for latitude and longitude

Two-way permutational ANOVA was used to test the significance of longitude and latitude of accidents on the number of crashes by binning the crashes into groups based on their spatial component. The model can be defined as follows:

$$\text{number\_of\_crashes}_{i,j,k} = \mu + \text{longitude}_i + \text{latitude}_j + \text{interaction}_{i,j} + \varepsilon_{i,j,k}$$

We start by studying the significance of the interaction of longitude and latitude. So we tested the following hypothesis:

$$\boldsymbol{H_0} : \text{interaction}_{i,j} = 0 \quad vs \quad \boldsymbol{H_1} : \text{interaction}_{i,j} \neq 0$$

After permuting the residuals under $\boldsymbol{H_0}$ and computing permutational distribution, we obtained the following histogram of test statistic and empirical distribution function:

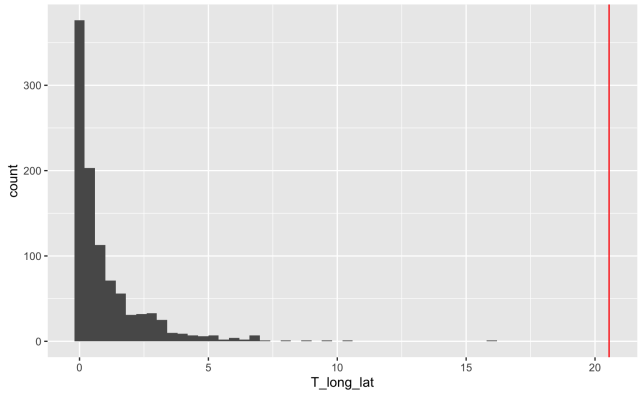

Figure 21: Histogram of Test Statistic

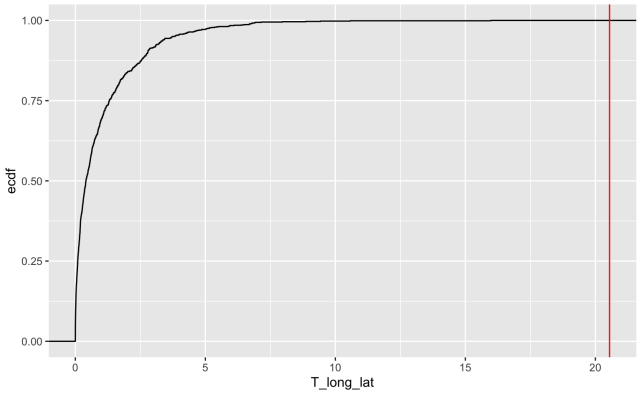

Figure 22: Empirical CDF of Test Statistic

The obtained p-value of the test was equal to 0, hence we conclude that the interaction of longitude and latitude of is significant. As a consequence we can conclude that the main effects, i.e longitude and latitude, independently are significant as well. The test to check the significance of main effects without interactions was omitted due to this fact.

3.2.5. GAM using latitude and longitude

A second possible approach to incorporate the spatial information in the model is by using the latitude ad longitude of the crashes, this approach is informed by the tests conducted in Section 3.2.4.
To do so we divided the map of the UK into a grid of 0.1 degrees of latitude and longitude ad tried to predict the number of daily crashes in each square as a function of the spatial components, the time and the weather conditions in a similar fashion to Section 3.2.3 by considering the most common weather pattern for that given day. The model we used is the following:

$$\begin{aligned}\log(\text{mean n}^\circ \text{ of collisions}_i) \sim \text{day of the week}_i + \text{wind}_i + \text{rainfall}_i + \text{year}_i + \text{Covid year}_i \\ + f_1(\text{day of the year}_i) + f_2(\text{latitude}_i, \text{longitude}_i)\end{aligned}$$

for the function f_1 we used a cubic spline basis, while for f_2 we found the best model to be a Gaussian process model with covariance function of the form:

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

and we set $\rho = 0.5$ degrees.
The fitted component can be seen if Figure 23, and it is clear that the we are capturing the distribution of the population of the UK since we have local maxima in the positions corresponding to the major cities in the UK, plotted in Figure 24. The dependence on the time day of the year was very similar to the one in Figure 14, and the coefficients were similar to the ones in Table 5, highlighting a negative impact of adverse weather conditions and a decrease as the years go on.
This model performed better both in terms of AIC and of explained deviance reaching 46.9%.

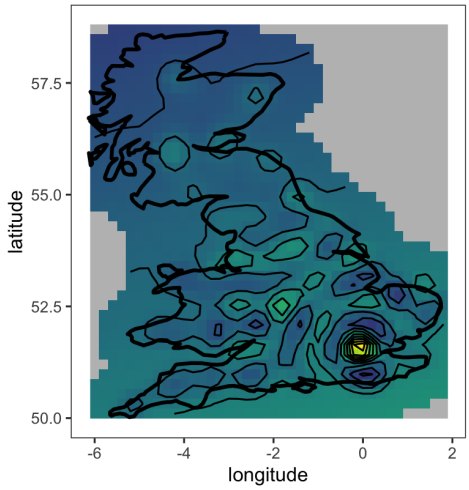

Figure 23: fitted latitude and longitude component

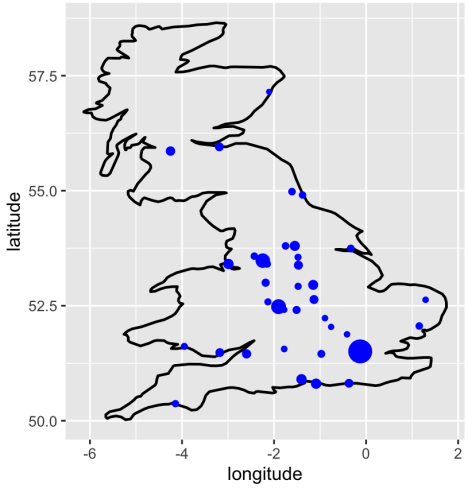

Figure 24: cities with a population larger than 200.000

3.2.6. Combining the previous two approaches

The next step is to combine the approaches outlined in Sections 3.2.3 and 3.2.5 in a single Generative Additive Mixed Model. The idea in this case it to use the a random intercept to fit the population of a given district and to use the latitude and longitude information to get more detailed information on the location. The final model is:

$$\begin{aligned}\log(\text{mean n}^\circ \text{ of collisions}_i) \sim \text{day of the week}_i + \text{wind}_i + \text{rainfall}_i + \text{year}_i + \text{Covid year}_i \\ + f_1(\text{day of the year}_i) + f_2(\text{latitude}_i, \text{longitude}_i) \\ + \text{district}_i\end{aligned}$$

We assumed the mixed effect to be normally distributed with zero mean, and we assumed a Poisson family for the model as in the previous cases. This significantly increased the performance of the model reaching an explained deviance of 62.3% and the highest AIC of any of the models. The resulting coefficient can be seen in Figures 25 and 26.

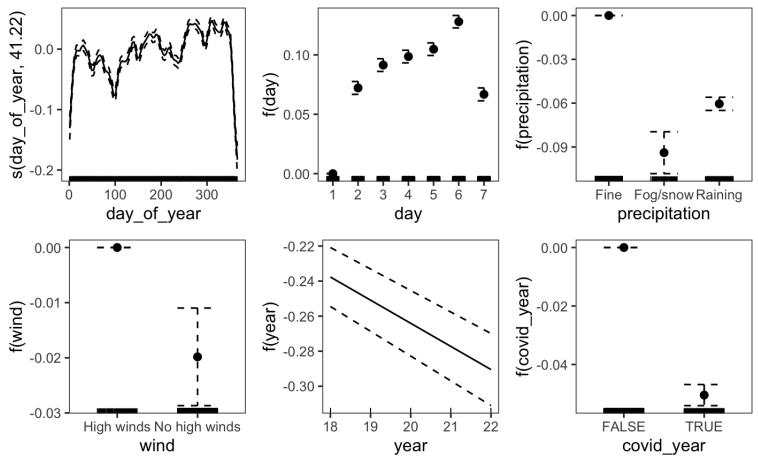

Figure 25: fitted values for the coefficient

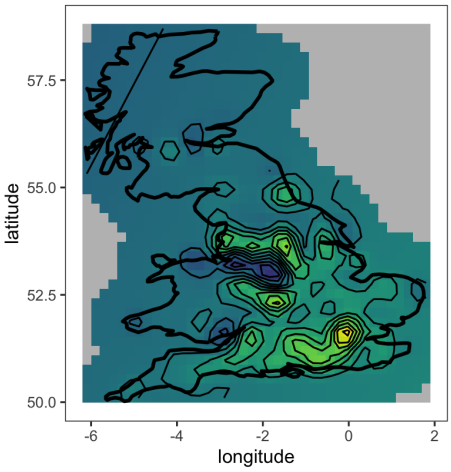

Figure 26: fitted spatial component

The model is able to model all of the features observed in Section 3.1. The temporal dependence in the year is modeled by the nonparametric function f_1, the dip during the pandemic by the dummy variable and the decrease in the years thanks to the parametric coefficient with negative sign. The effects of the spatial component is modeled by both f_2 and the random effects, that capture the distribution of the populations in both the spatial domain and in the different districts. The effects of the weather conditions is captured via the dummy variables related to the precipitation and wind.

4. Conclusions

The results of this analysis provide valuable information regarding the severity and the occurrence of road traffic accidents to the first responders, which in turn can be inform their future planning to provide a better service to the public.
Thanks to the analysis on the severity of the accident we can provide information on the time of the day and the type of roads that result in a more serious crash, allowing the emergency services to better plan their shifts and disposition accordingly.
Following the analysis on the daily number of collisions we can obtain valuable information on the number of first responders that should be on duty since they should be proportional to the random effects of the given districts, and the placement within the district should follow the nonparametric part of the model, we modeled also the dependence on the weather but it is not easily controllable.
We would like to highlight that all of the models work on the number of collision, a more principled approach would be to focus on the number of crashes per number of circulating vehicles in each year to have a more clear understanding for other stakeholders such as urban planners.

References

[1] Jacopo Diquigiovanni, Matteo Fontana, and Simone Vantini. The importance of being a band: Finite-sample exact distribution-free prediction sets for functional data, 2021.

[2] Department for Transport. Road safety data.

[3] R Lovelace, M Morgan, L Hama, and M Padgham. stats19: A package for working with open road crash data. *Journal of Open Source Software*, 4(33):1181, 2019.

[4] Martin Maechler, Peter Rousseeuw, Christophe Croux, Valentin Todorov, Andreas Ruckstuhl, Matias Salibian-Barrera, Tobias Verbeke, Manuel Koller, Eduardo L. T. Conceicao, and Maria Anna di Palma. *robustbase: Basic Robust Statistics*, 2024. R package version 0.99-2.

[5] Ida Ruts Peter J. Rousseeuw and John W. Tukey. The bagplot: A bivariate boxplot. *The American Statistician*, 53(4):382–387, 1999.

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

[7] Alessia Pini and Simone Vantini. *fdatest: Interval Wise Testing for Functional Data*, 2017. R package version 2.1.0.

[8] Alessia Pini, Simone Vantini, Bianca Maria Colosimo, and Marco 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, 03 2017.

[9] Laura M. Sangalli, Piercesare Secchi, Simone Vantini, and Valeria Vitelli. k-mean alignment for curve clustering. *Computational Statistics Data Analysis*, 54(5):1219–1233, 2010.

[10] Aymeric Stamm. *fdacluster: Joint Clustering and Alignment of Functional Data*, 2023. R package version 0.2.2.

[11] Simon N. Wood. *ocat: GAM ordered categorical family*, 2023. R package version 1.9-1.