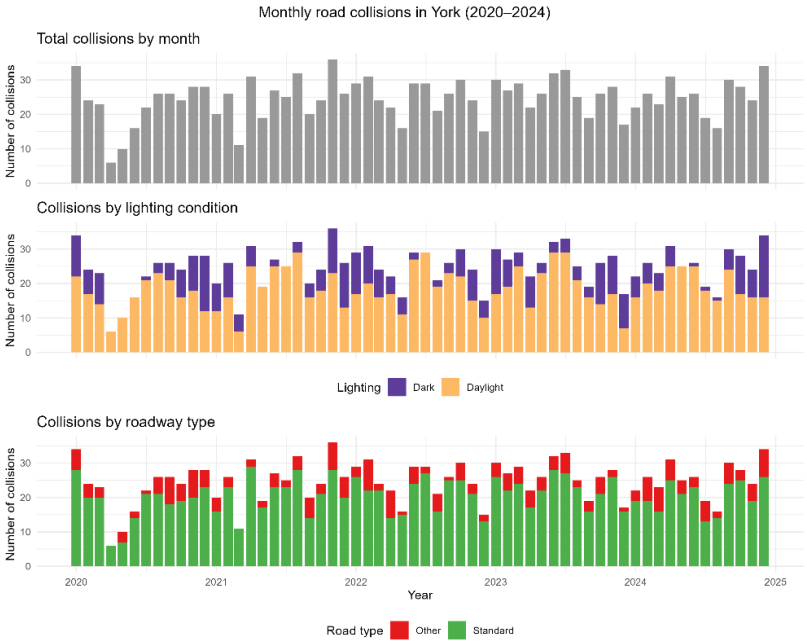
Modeling Spatial and Temporal Patterns in Road Crashes: Evidence from York, UK

1. Introduction

This project seeks to answer the **research question** *How do roadway design, environmental conditions, and time-of-year influence crash frequency in York, UK, after accounting for overdispersion and seasonal variation?*



This project focuses on the York 2023 subset (314 crashes) and an expanded York 2020-2024 sample (1,478 crashes).

Figure 1 shows monthly collision counts in York from 2020–2024, highlighting clear temporal variation as well as differences by lighting conditions and roadway type. These descriptive patterns motivate the use of a regression model that adjusts for seasonality, roadway structure, and environmental conditions.

2. Methodology

Spatial Aggregation: Collision locations are aggregated to a regular hexagonal grid with a fixed cell size of ~2 km. This resolution was chosen to preserve spatial heterogeneity and avoid excessive zero counts found at finer resolutions. We retained hexagons intersecting at least one collision to serve as analytic units, not causal locations. The raw dataset contains 314 crashes in 2023 and 1,478 crashes from 2020–2024, but the regression models are fit to a balanced panel of hexagon–month observations, including months with zero crashes.

Temporal Aggregation: In the 2023 analysis, we control time variation using month fixed effects to capture seasonal variation in crash frequency and month-specific shocks within the year. In the multi-year analysis, temporal effects are decomposed into month-of-year fixed effects to capture recurring seasonal patterns and year fixed effects to capture annual shifts (e.g., pandemic related changes). This structure is to avoid overfitting that would result from month-year indicators.

Covariate Construction: Roadway characteristics are summarized using a categorical variable where each for hexagon-month the roadway category is defined as the modal road type among crashes for that unit with hexagons with no crashes being assigned the baseline roadway category. Lighting and weather are summarized as binary indicators reflecting where any crash in that hexagon month occurred under dark lighting conditions or adverse weather conditions. These variables capture contextual associations with higher crash volumes rather than per-crash risk. Because environmental indicators are defined as whether *any* crash in a hexagon–month occurred under that condition, these coefficients reflect contextual association with higher crash volume rather than per-crash risk.

Statistical Model: Let Y_{it} denote the number of reported collisions in hex grid cell i during month t. Collision counts are modeled using a negative binomial regression with a log link and an offset to normalize the number of days in each month. The conditional mean $\mu_{it} = \mathbb{E}[Y_{it}]$ is specified as $\log(\mu_{it}) = \log(d_t) + \eta_{it}$ where d_t denotes the number of days in month t and η_{it} is a linear predictor.

For the 2023 analysis, η_{it} includes month fixed effects to capture seasonality along with roadway structure and environmental covariates:

$$\eta_{it} = \alpha + \gamma_t + \beta_{\boldsymbol{r}}^T \boldsymbol{r}_{it} + \beta_D D_{it} + \beta_W W_{it}$$

For the combined 2020-2024 analysis, temporal effects are decomposed into month-of-year and year fixed effects:

$$\eta_{it} = \alpha + \delta_{m(t)} + \lambda_{y(t)} + \beta_{\boldsymbol{r}}^T \boldsymbol{r}_{it} + \beta_D D_{it} + \beta_W W_{it}$$

Here $\boldsymbol{r}_{it}$ denotes roadway structure indicators, while D_{it} and W_{it} indicate whether any collision in a hexagon i during month t occurred under dark lighting or adverse weather conditions. Coefficients are reported on the log scale and exponentiated to obtain incidence rate ratios. Nested models are compared using the Akaike Information Criterion (AIC).

3. Results

This study compared nested models across two datasets using AIC. For both the 2023 analysis and the multi-year analysis, the **full model**—which included all roadway, environmental, and temporal covariates—was preferred compared to the base and road-only specifications. For instance, in the multi-year set, the AIC dropped significantly from the base model (AIC=6368) to the full model (AIC=5294). Because the full model had the strongest AIC scores, this paper focuses on the coefficient results from these models for comparison.

In this model, the coefficient estimates tell us how a factor (like dark lighting or complex roads) changes the expected monthly crash count in a hex cell compared to a baseline. A positive coefficient means the factor is associated with a higher expected crash count, and we can convert that estimate into a sample rate multiplier (e.g., “The crash count is X times greater”).

Environmental conditions showed the strongest associations with higher crash frequency, with the coefficients showing a consistently higher risk in the multi-year model. The crash rate in hexagon-months that experienced crashes during **dark hours** was dramatically higher than in daylight conditions. The expected crash count was 9.5 times greater ($e^{2.259} \approx 9.57, p < 0.001$) Crash rates that occurred during **adverse weather** also increased, with the expected crash count being approximately 3.9 times greater than good weather conditions ($e^{1.371} \approx 3.94, p < 0.001$).

Roadway design also proved to be statistically significant and associated with elevated risk. Road segments categorized as **complex** (e.g., multiple junctions) had an expected crash count over 3.3 times greater than the standard road baseline ($e^{1.202} \approx 3.33, p < 0.001$). Similarly **divided roads** had an expected crash count approximately 2.5 times greater than the standard baseline ($e^{0.922} \approx 2.51, p < 0.001$).

Coefficient	Estimate	Std. Error	z Value	p-Value
2023 Dataset (n=314)				
road_catdivided	0.639	0.431	1.48	0.139
road_catcomplex	0.969	0.424	2.29	0.022
road_catother	-	-	-	-
any_dark	1.911	0.174	10.98	<0.001
any_badweather	1.255	0.194	6.48	<0.001
2020-2024 Dataset (n=1478)				
road_catdivided	0.922	0.207	4.45	<0.001
road_catcomplex	1.202	0.158	7.63	<0.001
road_catother	1.372	0.760	1.81	0.071
any_dark	2.259	0.091	24.83	<0.001
any_badweather	1.371	0.101	13.64	<0.001

Table 1. Negative binomial regression results for monthly crash counts by hexagon

4. Conclusion

This study successfully demonstrates that **darkness, bad weather, and complex road designs** are strongly linked to higher crash frequency in York. However, to correctly understand our findings, we must acknowledge their limitations. The model only shows strong links, not causes, omitting key details, like traffic volume, that weren't readily available in the dataset. Area size also matters. We grouped the crash data into large 2 km hexagons describing risk across this area level, and the findings might change at more granular scales. Lastly, our model treats each area independently and does not capture any 'spillover' effect where risk in one area might influence its neighbors.

Future work should focus on supplementing this data by gathering information on traffic volume or counts (to measure collision opportunity) and demographic information (to account for behavioral risk). Additionally, it should explore the varying types of collisions that may occur whether it be car-to-car or collisions with active transport users such as cyclists or pedestrians.

Attribution of Sources

The road collision data used in this project was obtained from the UK Department for Transport STATS19 dataset via the **stats19** R package.

ChatGPT was used as a tool to assist navigating the **stats19** package, helping load in the proper data, as well as clarifying some confusing aspects of the package documentation. Chat GPT was also used in figuring out the best way to construct a negative binomial model using R, with it suggesting **MASS** and assisting with interpreting and setting up my model. All data processing, modeling decisions, and substantial analysis code as well as written content were authored by me.