MDM2 Project Progress Summary

Active Travel and Darkness Analysis

Shavarsh Melikyan, Group 3

February 25, 2026

1 Project Aim

The objective of this project is to investigate whether darkness acts as a deterrent to non-motorised transport in Bristol. Specifically, we analyse pedestrian and cyclist counts from hourly traffic sensors to determine:

- Whether darkness reduces active travel.
- Whether the effect varies spatially across clusters.
- How weather interacts with darkness.
- Whether lighting interventions could potentially increase activity.

Our approach estimates conditional associations rather than strict causal effects, while controlling for time-of-day, seasonality and weather.

2 Data Construction

We first merged all hourly sensor CSV files into a single dataset. Each row represents:

- A sensor
- At a specific hour
- With pedestrian (ped), cyclist (cyc), and car counts

We then added:

- Datetime (UTC)
- Hour of day
- Day of week

- Month
- Latitude and longitude

Weather variables (temperature, wind speed, rainfall) were cleaned from MIDAS data and merged by timestamp. The final modelling dataset is:

`big_table_with_weather_and_rain_with_geo_clusters_v2.csv`

3 Darkness Classification

Darkness was computed using solar altitude based on geographic coordinates and timestamp. This approach is physically grounded and more accurate than simply defining darkness by hour.

We created a binary variable:

$$Dark = \begin{cases} 1 & \text{if below solar threshold} \\ 0 & \text{otherwise} \end{cases}$$

4 Spatial Clustering

Sensors were grouped into three geographic clusters based on longitude and latitude. After refining the clustering to remove outliers, we retained three main clusters representing distinct spatial zones within Bristol.

This allows us to test whether the impact of darkness differs by area.

5 Model Selection

Because pedestrian and cyclist counts are non-negative integers, we initially considered Poisson regression. However, dispersion tests showed:

$$\text{Variance} \gg \text{Mean}$$

This indicates overdispersion. Therefore, we used a Negative Binomial (NB) Generalised Linear Model with a log link.

6 Model Structure and Justification of Relationships

The Generalised Linear Model (GLM) we estimated assumes a log-linear relationship between the expected count and explanatory variables. Formally:

$$\log(E[Y]) = \beta_0 + \sum_k \beta_k X_k$$

where Y is the hourly count (pedestrians, cyclists, or active total).

This implies:

- Effects are multiplicative on the original scale.
- Coefficients represent percentage changes.
- The expected count must remain positive.

Core Variables Included

Hour (categorical)

We include hour-of-day as a categorical variable to capture commuting patterns and daily cycles. This avoids imposing a linear time assumption and allows for flexible daily structure.

Day of Week

Controls for weekday vs weekend behavioural differences.

Month

Captures seasonal effects such as winter vs summer daylight differences and behavioural variation.

Temperature

Included as a continuous variable. Warmer temperatures are expected to increase outdoor activity.

Wind Speed

Higher wind speed may reduce cycling and pedestrian comfort.

Rainfall

Rain directly affects outdoor travel and is expected to reduce counts.

Spatial Cluster

Different areas of Bristol have different land-use characteristics (e.g. residential vs commercial). Including cluster controls for structural spatial differences.

Darkness Variable

The variable `Dark` captures whether the hour occurs in darkness based on solar altitude.

This is central to the research question. The coefficient on Dark estimates the percentage change in expected counts under darkness, holding other variables constant.

Interaction Terms

We included:

- Dark × Cluster

Justification:

The deterrent effect of darkness may vary spatially. For example, commercial zones may behave differently than residential areas.

We initially attempted:

- Dark × Hour

However, this caused a singular matrix error due to strong collinearity. Darkness is highly correlated with hour and month (e.g., 02:00 is almost always dark in winter). This created near-perfect multicollinearity.

The interaction was removed to maintain model identifiability.

Why Negative Binomial Instead of Poisson

Poisson assumes:

$$\text{Var}(Y) = E[Y]$$

In our data:

$$\text{Var}(Y) \gg E[Y]$$

This overdispersion makes Poisson inappropriate. The Negative Binomial model allows:

$$\text{Var}(Y) = E[Y] + \alpha E[Y]^2$$

which fits the observed variance structure.

7 Autocorrelation and the "All Zeros" Result

When testing residual autocorrelation using the Ljung–Box test, we obtained p-values extremely close to 0 at multiple lags.

This initially appeared unusual because all p-values were reported as 0. However, this does not mean the statistic is zero. It means:

$$p < 0.001$$

i.e., extremely strong evidence against the null hypothesis of no autocorrelation.

This indicates that residuals are serially correlated across hours.

This is expected in transport data because:

- Counts follow daily cycles.
- Weather effects persist over consecutive hours.
- Behavioural patterns are temporally structured.

To address this, we implemented cluster-robust standard errors at the sensor level.

This corrects inference (standard errors and p-values) without altering coefficient estimates.

8 Interpretation of Model Coefficients

Because we use a log link:

$$\% \text{ change} = (e^{\beta} - 1) \times 100$$

For example, a coefficient of -0.69 corresponds to approximately:

$$(e^{-0.69} - 1) \times 100 \approx -50\%$$

Thus, negative coefficients on Dark represent proportional reductions in expected counts during darkness.

9 Cluster-Robust Standard Errors

Residual diagnostics revealed strong autocorrelation:

- Significant lag-1 correlation
- Clear daily cycle (lag 24 spike)
- Ljung-Box test p-values ≈ 0

This indicates serial dependence within sensors. To account for this, we used cluster-robust standard errors at the sensor level.

10 Pedestrian Model

Dependent variable: `ped`

Controls:

- Hour (categorical)
- Day of week
- Month
- Temperature
- Wind speed
- Rainfall
- Spatial cluster
- Dark × cluster interaction

Results

Darkness is associated with a 40–50% reduction in pedestrian counts across clusters.

Graph Interpretation

Darkness Effect Bar Chart

- X-axis: Cluster number
- Y-axis: Percentage change in expected pedestrian count

Negative values indicate reduction under darkness.

Residuals vs Fitted Plot

- X-axis: Predicted count
- Y-axis: Pearson residual

The funnel shape is expected in count models, as variance increases with the mean.

ACF Plot

- X-axis: Lag (hours)
- Y-axis: Autocorrelation

Spike at lag 24 reflects daily commuting cycles.

11 Cyclist Model

Dependent variable: `cyc`

Cyclist responses were more heterogeneous. Some clusters showed smaller reductions compared to pedestrians, likely reflecting commuting behaviour.

12 Combined Model (Pedestrians + Cyclists)

Dependent variable:

$$active = ped + cyc$$

Estimated darkness effects:

- Cluster 0: $\approx -30\%$
- Cluster 1: $\approx -32\%$
- Cluster 2: $\approx -50\%$

Cluster 2 appears most sensitive to darkness.

13 Specification Adjustments

We attempted to include a Dark $\times$ Hour interaction. However, this caused a singular matrix error due to strong collinearity between darkness and hour. We removed this interaction and retained:

- Dark $\times$ Cluster
- Dark $\times$ Rain

The simplified model converged correctly.

14 Conclusions So Far

- Darkness is strongly associated with reduced pedestrian activity.
- Cyclist effects are heterogeneous.
- Combined active travel drops significantly, especially in Cluster 2.
- Weather effects behave as expected (rain and wind negative).

While we estimate conditional associations rather than strict causality, results suggest that lighting interventions may be most beneficial in Cluster 2.

15 Next Steps

To inform policy decisions, we plan to:

- Simulate counterfactual predictions with Dark = 0.
- Rank sensors by predicted uplift.
- Construct cumulative benefit curves to assess diminishing returns.

This will allow us to recommend targeted lighting rather than blanket installation.

Nota Bene: Clarification on Diagnostic Graphs and Autocorrelation

This section provides additional clarification on the diagnostic graphs produced during the modelling process, including how they were constructed and how they should be interpreted.

Pearson Residuals vs Fitted Values

This plot was generated using model output and standard diagnostic tools in `statsmodels`. The residuals were computed automatically by the GLM framework.

X-axis: Fitted (predicted) values from the model, i.e. the expected count $\hat{\mu}_i$ for each observation.

Y-axis: Pearson residuals,

$$r_i = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{V}_i}}$$

where Y_i is the observed count and $\hat{V}_i$ is the model-implied variance.

This graph checks whether the model's mean structure is appropriate. In count models, a "funnel" shape is expected because variance increases with the mean. Therefore, widening spread at higher fitted values is normal and not necessarily a model flaw.

Autocorrelation Function (ACF)

The ACF plots were generated using the built-in `plot_acf()` function from `statsmodels`. The number of lags (e.g., 50) was explicitly defined in the function call.

Important clarification: Specifying `lags=50` does *not* mean only 50 hours of data were used. Instead, it means that we examined correlation up to 50 hours back in time.

For each lag k, the function computes:

$$\rho(k) = \text{Corr}(r_t, r_{t-k})$$

using almost the entire year's data. For example, at lag 50, the model compares residuals at hour t with residuals 50 hours earlier for all valid time pairs across the full dataset.

X-axis: Lag in hours (how many hours back in time).

Y-axis: Autocorrelation coefficient, bounded between -1 and 1.

Lag 0 is always equal to 1 because a variable is perfectly correlated with itself. This value is not interpreted.

Spikes at lag 1 indicate short-term persistence (residual at hour t depends on hour $t-1$). Spikes at lag 24 reflect daily behavioural cycles, as transport patterns repeat approximately every 24 hours.

This behaviour is expected in hourly transport data and does not indicate a computation error.

Why Only One Sensor Was Used for ACF

Autocorrelation requires a continuous time series. Because the dataset contains multiple sensors (panel data), we selected a single representative sensor's residuals for diagnostic inspection. This was a methodological choice to illustrate serial dependence, not to imply that the result is sensor-specific.

Ljung–Box Test and "Zero" p-values

The Ljung–Box test was applied to formally test for residual autocorrelation.

In several cases, p-values were reported as 0.0. This does *not* mean the statistic was zero. It means:

$$p < 0.001$$

i.e., extremely strong evidence against the null hypothesis of no autocorrelation.

This confirmed the presence of serial dependence, which justified the use of cluster-robust standard errors at the sensor level.

Darkness Effect Plot

The darkness effect plots display estimated percentage changes derived from model coefficients.

Because the model uses a log link, percentage change is computed as:

$$(e^{\beta} - 1) \times 100$$

X-axis: Geographic cluster. **Y-axis:** Estimated percentage change in expected counts due to darkness.

Negative values indicate reductions in activity during dark hours.

Implementation Notes

All diagnostic plots were generated using standard Python statistical libraries:

- `statsmodels` for GLM fitting and residuals
- `statsmodels.graphics.tsaplots` for ACF
- `matplotlib` for plotting

The statistical formulas (residuals, correlation, percentage change transformation) are based on the GLM framework rather than manually defined custom equations.

Conclusion on Diagnostics

The observed autocorrelation patterns are expected in high-frequency transport data. They reflect daily behavioural cycles and short-term persistence rather than model misspecification.

The diagnostic results therefore justified the methodological choice of cluster-robust standard errors, while retaining the Negative Binomial GLM structure.