FUTURE-DATA
Air Quality and Urban Trees: A Data-Driven Assessment for Dublin City Council

Author:
Minh Trinh
Sania Suneeth
Loimar Vianna
Qin Guo

Introduction

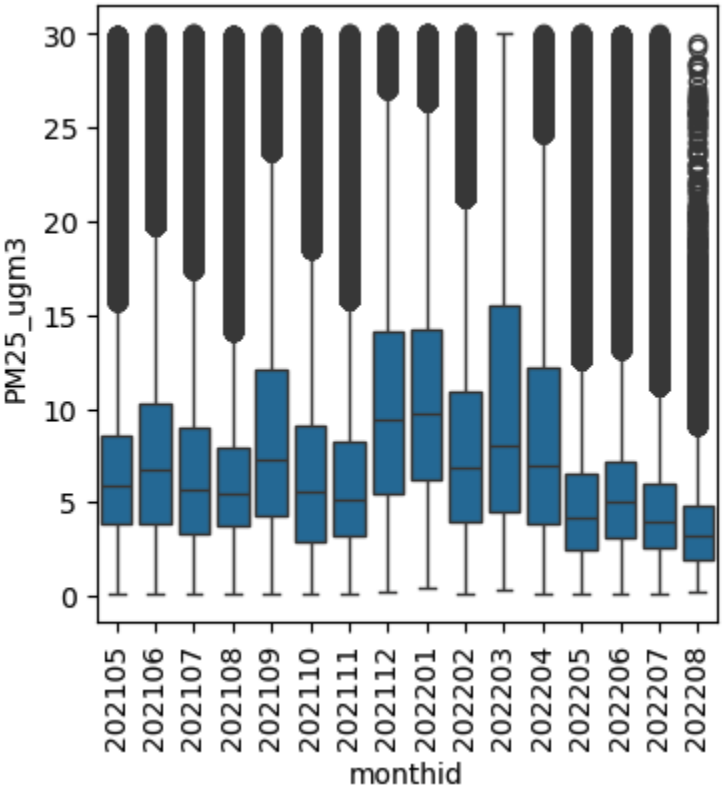
At the request of Dublin City Council (DCC) , FUTURE-DATA conducted a focused study to better understand the factors influencing air pollution in Dublin. Some of the key points of the analysis is:

+ We will analyse the relationship between tree coverage and the level of pollution, controlling for the volume of traffic in the Dublin area as a whole.
+ We will later divide Dublin into 5 areas for analysis as each area may have distinctive features that can affect the tree coverage and pollution relationship.
+ We will focus on Late Summer (represented by July and August 2021) and Winter months (represented by Dec 2021 and Jan 2022) which have a contrasting level of tree leaves density.
+ PM2.5 pollutants are the focus of this analysis as they are known to be captured and absorbed by vegetation.

Data examination

Our main dataset is the Google Air View dataset. Each day has a varying recording time range, so we restricted our analysis to the **10:00–17:00** window, which provides the most consistent overlap across the four months of interest. The data only includes weekdays. Missing data for the PM25_ugm3 variable was minimal — typically less than 5% of daily observations — and given the large sample size (300k–400k records/month) and our plan to aggregate data, we determined that removing missing values would not affect the results. As shown in Figure 1, PM2.5 data is right-skewed, with most values below 30 µg/m³, though extreme values reach 2000 µg/m³. The least polluted months are July and August 2021, while December 2021 and January 2022 show peak pollution. It will be interesting to see how tree coverage fair in these two contrasting seasons. The dataset also contains longitude and latitude of measure point which will help with mapping location with other dataset.

Figure 1: Monthly distribution of level of PM2.5 in ug/m3 using measure point



Next for the road traffic date, we will be using data from the DCC's SCATS dataset (The Sydney Coordinated Adaptive Traffic System) which comes from sensors at junctions to monitor traffic volume on an hourly basis. The system is designed to adapt traffic light sequences based on real-time vehicle flow. We processed data from the same four months and retained only weekday data between 10:00–17:00. We used the total_volume column to measure traffic intensity for our analysis. There is no missing data found in the variable we are interested in. We summed up the data to have the total level of traffic at each junction for the 4 months of interest. One caveat is that the data have no Longitude and Latitude. We have to acquire a dimension table from the DCC website to do the mapping using the Siteid column. After mapping, about 289 out of 905 of the unique SiteID can not be mapped with any coordinate. We will have to remove them. There is also a Region column, which we will also rely on later to segment Dublin into different areas: City centre (CCITY) , North City (NCITY), South City (SCITY), West City (WCITY1) and Dock Area (DCC1). After some

inspection, we have also found a number of junctions with 0 traffic at any time. We have personally inspected the location of some of the junctions using the mapped coordination and found out they are in the middle of normal streets. Our judgment is that this is an error in the data and we have decided to remove them. After removing missing data and cleaning, the number of aggregated junction observations reduced from 3,533 to 2,155.

We obtained tree data from Mapping Green Dublin, recorded 2017–2020. Although not perfectly aligned with our 2021–2022 time frame, this was the closest available. The dataset includes over 300k individual tree records in Dublin with coordinates, but attributes like species and height had high levels of missingness. Fortunately, all location data was intact.

Overall, we have found all the dataset required for our analysis, with coordination for mapping. Google's airview dataset has minor missing data issues but provides a comprehensive coverage of Dublin street at coordinate level with many kinds of pollutants recorded. Due to the scope of this research we can only focus on PM2.5. We also wish the data covered a longer period of time so it matches our tree data time range.SCATS's missing data issue was major, reducing a large number of observations. Tree data is sufficient for our use, but there can be so much potential with tree's attribute data if there is less missingness.

Data merging

We merged the three datasets using coordinates. After identifying the overlapping area between them, we defined a common analysis boundary shown in the grey area in Figure 2 below. The colored circle in the figure is junction observation with Regions as different colors. The original regions from SCATS mapping files sometimes do not reflect the true geolocation. We have tried to manually fix most of that but some mixup still remains.

Figure 2: Area of analysis and data point of junction with Region as colour mapping

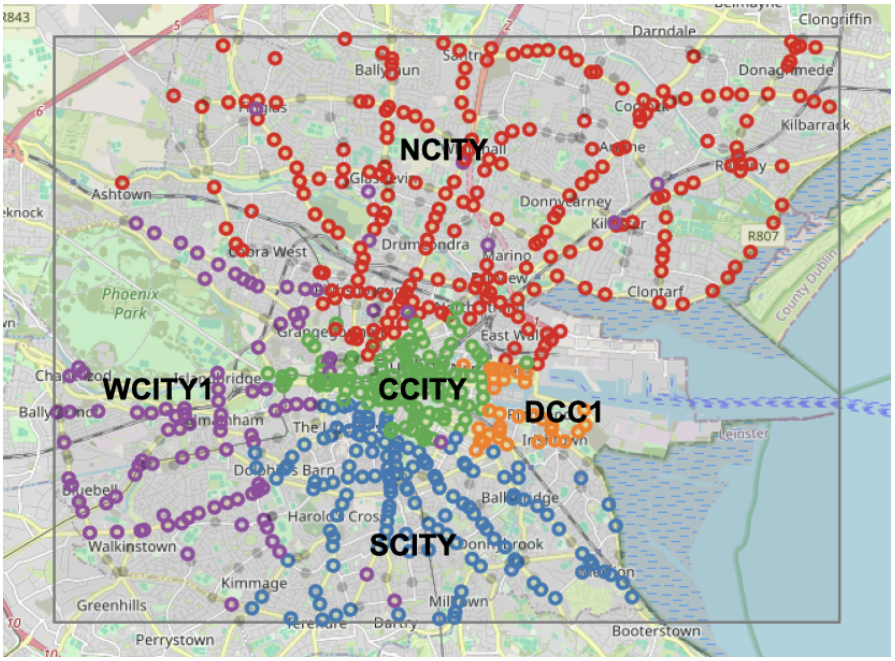


Figure 3: 100m radius circle from junction covering trees and pollution points

We defined circular zones around each junction (e.g., 50m or 100m radius). Within each circle, we calculated the three monthly figures: the average PM2.5 level, number of trees (always stay the same), and total traffic volume. An illustration is shown on Figure 3 above. Dark red dots represent circles with a 100m radius range. Green dots are trees and purple dots are pollution observations with darker shades being higher in PM2.5 concentration.

The final merged data will have the form:

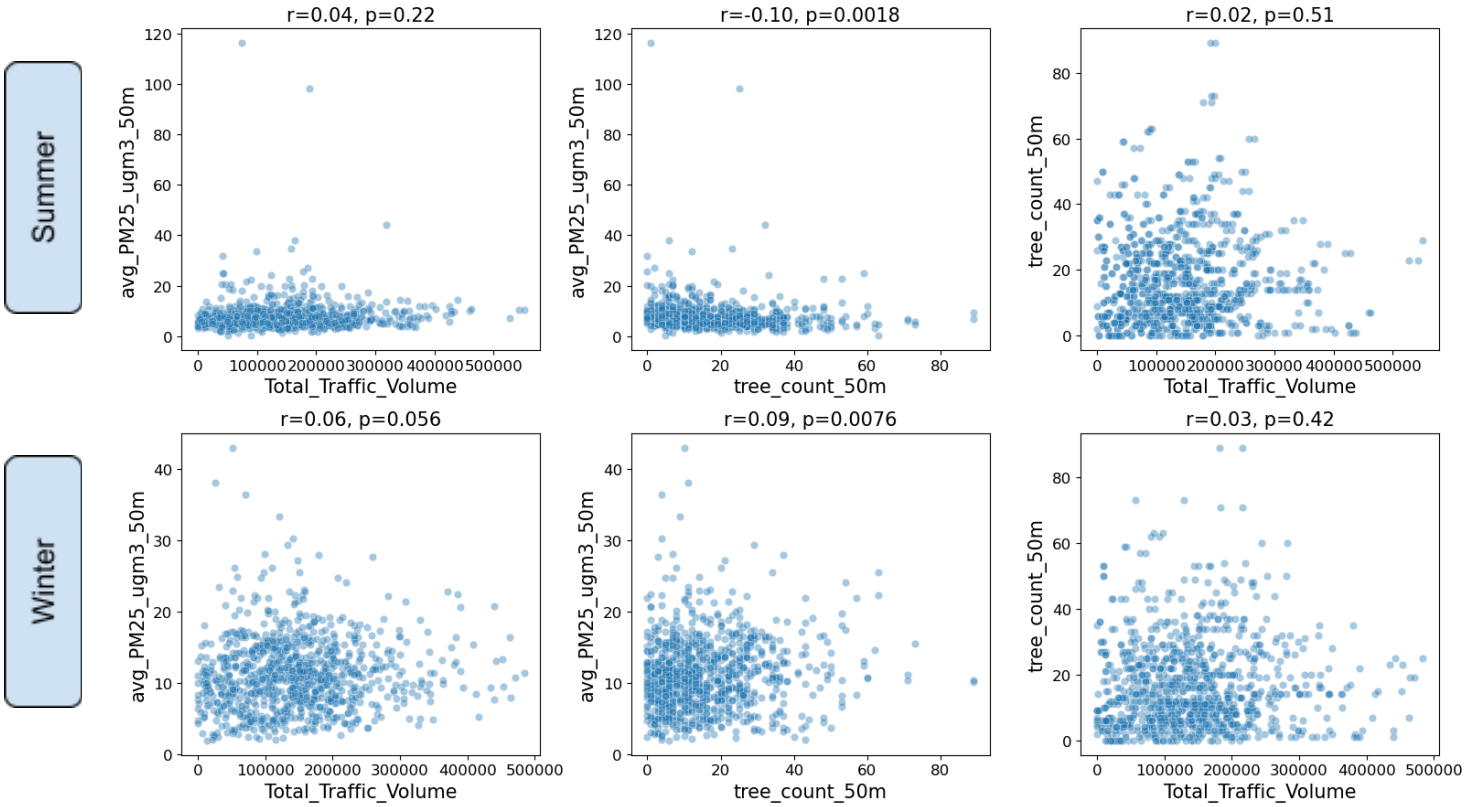
JunctionID	Latitude	Longitude	Monthid	Total_traffic_vol	no_tree_in_x_m	avg_pollution_in_x_m

After applying the common boundary so that we only use junctions within the border, we have 1989 observations left with distribution: NCITY: 729, CCITY: 446, SCITY: 407, WCITY1: 296, DCC1: 101

Analysis

We compared results for 50m vs 100m radius zones. The 50m radius provided more explanatory power, so we used it for all subsequent analysis. We will have a look at the pairwise relationship between our variables of interest by late summer season and winter season, with correlation and its p-value in Figure 4

Figure 4: Pairwise relationship between variables



Now we see signs that total traffic volume are correlated with pollution in both summer and winter times (I am sure p-value will be significant after some outlier removal). Total traffic volume and count of trees are not correlated so we don't have to worry about multicollinearity in later regression. One surprising finding is that tree counts are negatively correlated with pollution in summer but positively correlated with pollution in winter. After doing some research in the literature, we have found that there are two ways trees can affect air pollution. **Absorption effect**: tree leaves surfaces are effective in capturing PM2.5. **Aerodynamic effects**: Tree reduces air ventilation, leading to increased pollution. In summer, tree leaves density is at its highest. The absorption effect overcomes aerodynamic effects so more trees mitigate pollution. While in winter lower tree leaves density reduce both aerodynamic and absorption effect, however aerodynamics dominates, making more trees worsen pollution. The multiple linear regression result in Table 5 with avg_PM25_ugm3_50m as dependent variable and Total_Traffic_Volume & tree_counts_50m as independent variables also supports finding from the plot and correlation. Also We have removed any observations with measures of PM2.5 concentration > 20 (66 observations in total) as they tend not be in sync with the general trend. In the regression I also divide into two models summer and winter

Table 5: Regression results (excluding intercept)

Season	Variable	Coef	p_value	n_obs
Summer	Total_Traffic_Volume	0.000004	0.0006	930
	tree_count_50m	-0.045510	0.0000	930
Winter	Total_Traffic_Volume	0.000004	0.0049	877
	tree_count_50m	0.031886	0.0021	877

We also examined each region more closely by running separate multiple linear regressions for summer and winter in all five regions. We chose not to use a single regression model with a region variable, as we believe the slope may differ across regions, and including interaction terms would make the analysis unnecessarily complex. As a result, each regional model contains a relatively small number of observations. Given this, we adopt a less conservative threshold and interpret coefficients with p-values ≤ 0.1 as statistically meaningful at this stage

Table 6: Regression results by regions (p values <=0.1 only and excluding intercept)

Model ID	monthid	Region	Variable	Coef	p_value	n_obs
4	Summer	NCITY	Total_Traffic_Volume	0.000007	0.000011	346
16	Winter	CCITY	Total_Traffic_Volume	0.000009	0.001606	209
13	Summer	DCC1	Total_Traffic_Volume	0.000011	0.087109	40
7	Summer	WCITY1	Total_Traffic_Volume	0.000006	0.091471	140
5	Summer	NCITY	tree_count_50m	-0.027539	0.008526	346
11	Summer	SCITY	tree_count_50m	-0.033247	0.017554	195
20	Winter	NCITY	tree_count_50m	0.045206	0.020169	303

Here the level of traffic is found to have a positive impact on the summer month in the North, West, and Dock area but only have a positive impact on winter month in the City Center areas. Tree counts are only effective in mitigating PM2.5 in the North and South City during summer, while in winter, tree coverage is associated with worsened pollution in North City.

Recommendation

Overall, the implication of the results is that if we want to reduce pollution, it is a good idea to reduce the level of traffic in the North, West and Center of the City. South City level of pollution somehow doesn't have a relationship with the level of traffic in this research. This should call for more investigation. For the rest of the area, some only show relationships in winters, some only show in the summer. This should also require closer examination. For future research I think it is a much better idea to examine the effect in a more granular scale like hour rather than month.

We usually have a dangerous misconception of planting trees to reduce pollution. This research has shown that it is not that simple. The perfect example is the North City area where tree counts help mitigate pollution in the summer but worsen pollution in the winter. Our speculation is that there is a great mix of seasonal trees in the area where leaves are shed in the winter. While in the South City trees can mitigate pollution in summer without having any effect in the winter. Again our speculation is that there are a lot of evergreen trees in this area. There could be other factors to consider but this research shows signs that we should try to plant more evergreen trees instead of seasonal trees.

This research enables us to discover which direction to further investigate. There are alot of limitations in the research scope because it only revolves around the junctions area. Besides, there are other variables that can have a major impact on the level of pollution like size of trees, tree species, temperatures, wind speed, and other human activities. Later research should include them in the analysis to have a more complete picture of air pollutants contributors.