AIR QUALITY INSIGHTS FROM A DENSE SENSOR NETWORK AND HIGH-RESOLUTION LAND-USE IN CHICAGO

GRANT COLEMAN

LADCO SUMMER INTERN

JULY 2026

PREPARED FOR THE LAKE MICHIGAN AIR DIRECTORS CONSORTIUM (LADCO)

1 Introduction

Nitrogen dioxide (NO_2) is both a directly harmful air pollutant and a key precursor to ground-level ozone, a secondary pollutant that drives much of the health burden associated with urban air quality in the United States. Long-term NO_2 exposure has been linked to an estimated 170,850 premature deaths per year nationally, with the largest burdens falling on residents of major metropolitan areas (Camilleri et al., 2023). In the Chicago metropolitan area, ozone remains a persistent regulatory challenge, and reducing oxides of nitrogen oxides (NO_x) emissions is central to regional strategies for bringing ozone concentrations into compliance with federal standards. Effective NO_x control, however, depends on first identifying which emission sources and environments are actually driving elevated NO_2 concentrations across the urban landscape, information that has historically been difficult to obtain at fine spatial resolution. Although these relationships are well understood at broad regional scales, they remain considerably more difficult to characterise at the neighbourhood scale, where emission sources often vary substantially over short distances.

Regulatory monitoring networks, which rely on a small number of reference-grade instruments, are designed to characterise compliance with air quality standards at the regional scale rather than to resolve block- or neighbourhood-level variation in pollutant concentrations. This coarse spatial resolution limits their ability to attribute NO_2 patterns to specific nearby sources such as major roadways, rail yards, or industrial facilities. Satellite-based approaches, including those using TROPOMI observations, have improved understanding of urban-scale NO_x emission patterns but remain constrained by pixel sizes that are large relative to the scale of individual emission sources in dense urban environments (Goldberg et al., 2024). Prior work using such networks to observe near-surface pollutant sources in urban settings has demonstrated their value by identifying spatial heterogeneity that current regulatory networks cannot capture (Shusterman et al., 2018), and hourly monitoring data have similarly been used to derive city-level NO_x emission signals in Chicago specifically (de Foy, 2018), though most applications to date have focused on pollutants other than NO_2.

Dense low-cost sensor networks complement these approaches by trading some measurement precision for dramatically increased spatial coverage, allowing pollutant variability to be examined at scales inaccessible to either regulatory monitors or satellite retrievals.

The Open Air Chicago network was developed via a partnership involving the City of Chicago and the University of Illinois Chicago and has deployed a dense array of low-cost NO_2 and fine particulate matter ($PM_{2.5}$) sensors across the city. The sensors are sited on a nominal 1.5-km grid (0.87-0.93 miles between adjacent sensors), designed following U.S. EPA Clean Air Act network design criteria and the Chicago Environmental Justice Index (City of Chicago, 2026). This network has generated a multi-month record of concentrations at the densest spatial resolution in the country (City of Chicago, 2026). This dataset creates an opportunity to move beyond simply mapping average concentrations and instead ask a more mechanistic question: do sensors that share similar temporal signatures, daily, weekly, and seasonal patterns of NO_2, also share similar built-environment characteristics, such as proximity to major roadways, rail infrastructure, or industrial land use? If so, clustering of sensors based on their NO2 patterns can provide a „fingerprint" of the emission sources that drive that NO2 distribution, offering a data-driven complement to conventional emissions inventories and providing actionable spatial information for NO_x control strategy design.

A substantial amount of literature has examined the relationship between traffic, industrial activity, and urban air pollution, including work linking vehicle emissions to spatial and equity disparities in exposure (Lang et al., 2025) and studies quantifying pollution impacts from specific land uses such as warehousing, which has expanded rapidly in the Chicago region and elsewhere (Kerr et al., 2024). Fewer studies, however, have examined whether temporal patterns derived from dense sensor data correspond with systematic land use and infrastructure patterns.

Recognising the value that could be found in combining such a dense sensor network with a high-resolution land use dataset, we at LADCO were guided by the motivating question: what emission sources drive the distribution and dynamics of NO_2 concentrations in Chicago? While that is not the question that this study sets out to answer, it will be the basis for future analysis. During this study, we set out to answer: are there associations between land use and Open Air Chicago NO_2 concentrations? This study addresses that by using nine months (September 2025-May 2026) of NO_2 data from the Open Air Chicago sensor network. We first applied principal component analysis (PCA) and hierarchical clustering to derived temporal features (percentile-based summaries of NO_2 concentration by month and by day-of-week/time-block combinations) to group sensors according to shared temporal concentration behaviour, independent of any land-use information. We then characterised the built environment surrounding each sensor using land-use, roadway, and rail data across three buffer radii (0.5, 1, and 2 km), and tested whether sensors assigned to the same temporal cluster also share distinctive land-use, traffic, or rail-proximity profiles. Specifically, we asked: (1) how many distinct clusters of sensors exist across the sensor network, and how do they differ in magnitude, diurnal shape, and seasonality of their NO2 concentrations, (2) are these clusters associated with different land-use compositions, roadway traffic volumes, or rail amount, (3) which categories of land use and infrastructure are most strongly associated with the highest-concentration clusters, and are these associations robust across buffer scales, and (4) are these observed associations consistent across multiple spatial scales, or are they sensitive to the buffer distance used to characterise the surrounding built environment?

The remainder of this paper is organised as follows. Section 2 describes the Open Air Chicago sensor data and the CMAP land-use and TIGER/Line roadway and rail datasets used to characterise each sensor's surroundings. Section 3 details the temporal feature engineering, dimensionality reduction, and hierarchical clustering procedures, along with the buffer-based land use characterisation methodology. Section 4 presents the resulting sensor clusters, their temporal signatures, and their associated land use and infrastructure profiles, interpreting these patterns in the context of Chicago's emission sources. Section 5 concludes with implications for NO_x control strategy and directions for future work linking sensor-derived clusters to formal emissions inventories.

2 Data

2.1 Study Area

The study area is the City of Chicago and the surrounding portions of Cook County, Illinois, corresponding to the deployment of the Open Air Chicago sensor network. This area encompasses a heterogeneous mix of dense urban residential neighbourhoods, commercial corridors, heavy-industrial districts, and major intermodal freight infrastructure, making it well suited to a land use-informed analysis of NO_2 spatial variability.

2.2 Open Air NO_2 Concentration Data

The primary dataset is NO_2 concentration data from the Open Air Chicago sensor network, a dense, low-cost air-monitoring deployment operated in partnership with the City of Chicago Department of Public Health and the University of Illinois Chicago. Data were obtained via the City of Chicago's Socrata open-data portal. The analysis aggregated the reported sub-hourly NO_2 concentrations to hourly concentrations, in parts per billion (ppb), for each sensor over the study period of September 2025 through May 2026. Each observation is associated with a sensor identifier (*sensor_name*), a timestamp (*date*), and geographic coordinates (*latitude, longitude*). There were a total of 7,153,586 observations gathered by the 277 sensors over the nine months. Data completeness, defined as the percentage of hours in a given month with at least one reported observation, varied both by sensor and by season (Figure 1); completeness was generally high across the network but declined at a subset of sensors during the winter months, when cold temperatures and reduced solar charging are more likely to interrupt reporting from the network's solar-powered units. This seasonal pattern is a direct motivation for the completeness-based filtering described in Section 3.2 and is revisited as a limitation in Section 4.5.

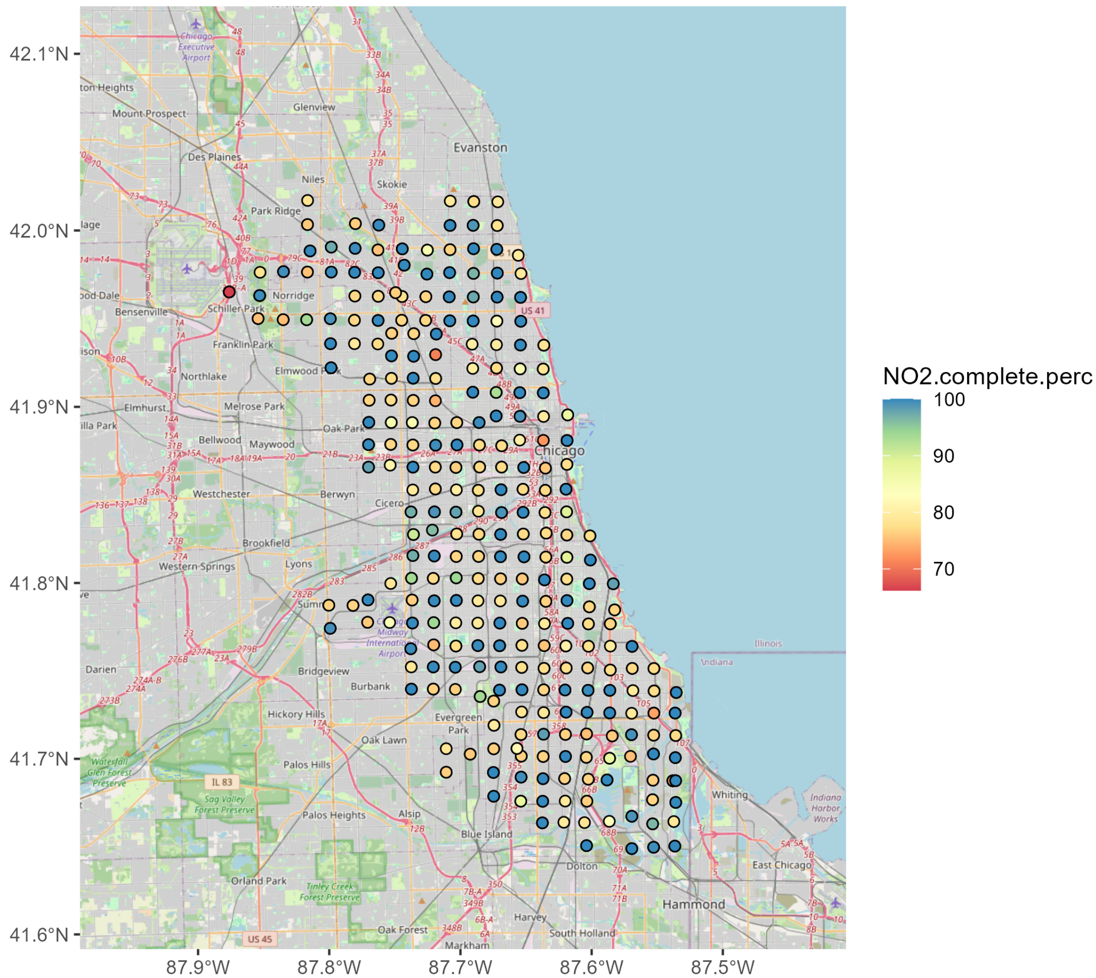

Figure 1. Percentage of hours per month with at least one reported NO_2 observation, by sensor. Completeness is defined on an hourly basis, so that any observation within an hour counts as a complete hour. Sensors are mapped at their deployed locations across the Open Air Chicago network

2.3 Land Use and Infrastructure Data

Three geospatial datasets were used to characterise the built environment surrounding each sensor:

Land use. The Chicago Metropolitan Agency for Planning (CMAP) 2023 Land Use Inventory for Northeastern Illinois is a parcel-level land-use classification covering the region at fine spatial resolution. Each parcel is assigned a numeric land-use code (e.g., 1420 = "General Industrial," 1511 = "Rail ROW," 6000 = "Non-Parcel Area") following CMAP's standard taxonomy, which we use to derive the dominant land use categories surrounding each sensor.

Roadways. Illinois Department of Transportation (IDOT) Annual Average Daily Traffic (AADT) data provided traffic volume (vehicles/day) by road segment, joined to the U.S. Census Bureau TIGER/Line road network. Road segments are classified by functional class (e.g., S1100 = primary roads/interstates and expressways, S1200 = secondary roads/arterials and state highways, S1400 = local roads), which we use to construct separate volume and length metrics by road type.

Rail. U.S. Census TIGER/Line rail feature data was used to quantify the length and amount of unique rails of rail infrastructure within each buffer zone.

Two additional reference layers were used solely to support geoprocessing rather than as analytical variables: the Illinois State Geological Survey's Illinois County Boundaries (polygon and line) dataset, used to constrain and clip spatial operations to the appropriate administrative extent, and Cook County Government's Lakes and Rivers hydrographic polygon dataset, used to exclude open-water area from land-use percentage calculations so that water bodies did not artificially deflate the share of developed land use within a buffer.

3 Methodology

3.1 Analytical Framework

The analysis proceeds in three stages. First, we constructed a set of temporal features for each sensor that summarised its NO_2 concentration behaviour across monthly and weekday/time-block dimensions, then used principal component analysis (PCA) to reduce this high-dimensional feature space while retaining the dominant modes of temporal variability. Second, we applied hierarchical clustering to the reduced (principal-component) representation to group sensors with similar temporal signatures, selecting the number of clusters using a combination of internal validity indices and dendrogram/silhouette diagnostics. Third, we characterised the land use, roadway traffic, and rail infrastructure surrounding each sensor at three buffer radii and tested whether these built-environment characteristics differ systematically across the clusters identified in the second stage. Clusters are defined purely from patterns in the NO_2 data, with no land use information used in their construction, so that any subsequent association between cluster membership and land use constitutes an independent validation of the clustering rather than a circular result.

All analyses were conducted in R using the tidyverse collection of packages for data manipulation, lubridate for time handling, openair for time-series preprocessing, FactoMineR and factoextra for PCA and its visualisation, cluster and NbClust for hierarchical clustering and cluster-number selection, and sf with ggspatial for spatial visualisation of cluster assignments. Geoprocessing of land use, roadway, and rail data was performed in ArcGIS Pro, primarily using the Buffer, Pairwise Intersect, and Tabulate Intersection tools, prior to import into R for statistical analysis.

3.2 Data Preprocessing and Quality Control

Timestamps were parsed and localised to the local America/Chicago time to ensure that diurnal and day-of-week patterns reflected local solar and social time rather than UTC. From each timestamp, we derived calendar month, weekday, and hour-of-day, and classified each hour into one of four time blocks (05:00–07:00, 08:00–10:00, 11:00–17:00, 18:00–04:00) intended to approximate morning commute, mid-morning transition, midday, and evening/overnight periods.

For each sensor, we computed the 10th, 50th (median), and 90th percentiles of NO_2 concentration within each month and within each weekday x time-block combination, along with the corresponding observation count. Using percentile summaries rather than raw hourly values reduces the influence of noise on the clustering procedure described in Section 3, while still preserving information about the shape (not just the central tendency) of each sensor's concentration distribution.

Sensors and monthly/day-of-week variables with substantial missing data were removed prior to analysis. Specifically, a completeness threshold of 50% was applied in both directions: temporal variables with fewer than 50% non-missing values across sensors were dropped, followed by sensors with fewer than 50% non-missing values across the retained variables. Any sensor–variable combinations still containing missing values after this filtering step were excluded via complete-case analysis prior to the principal component analysis, since PCA requires a complete data matrix. 268 sensors were retained after completeness filtering.

Because the distribution of NO_2 summary values was strongly right-skewed, all values entering the clustering procedure were log-transformed (using a log(1+x) transformation to accommodate zero values) prior to standardisation. This transformation was confirmed by visual inspection of histograms of the raw and log-transformed data.

Land-use codes in the sensor metadata table were mapped to descriptive category labels using a lookup table reflecting CMAP's published classification scheme (Section 2.3), enabling interpretable summaries of dominant land use by sensor and by cluster.

3.3 Temporal Feature Construction

To characterise each sensor's temporal NO_2 signature independent of noise in any single hourly observation, we constructed two complementary sets of percentile-based summary features:

Monthly features: the 10th, 50th, and 90th percentiles of NO_2 concentration for each sensor in each calendar month, capturing seasonal variation in both typical concentrations (median) and the spread of the distribution (10th/90th percentiles).

Weekday x time-block features: the same three percentiles computed for each sensor within each combination of day of week (Monday-Sunday) and one of four time blocks (05:00-07:00, 08:00-10:00, 11:00-17:00, 18:00-04:00), capturing diurnal and weekly patterns associated with, for example, commuting-related traffic peaks versus overnight background concentrations.

These two feature sets were reshaped from long to wide format and merged by sensor into a single sensor-by-feature matrix, in which each row represents one sensor and each column represents one percentile statistic for one month or one weekday/time-block combination. This yielded a matrix with 837 temporal variables per sensor before completeness filtering. This matrix was then filtered and log-transformed as described in Section 2.5, and standardised (zero mean, unit variance) column-wise before PCA, so that variables with intrinsically larger numeric ranges (e.g., 90th-percentile summer concentrations) did not disproportionately dominate the reduction relative to variables with smaller ranges (e.g., overnight medians).

3.4 Dimensionality Reduction via Principal Component Analysis

PCA was applied to the standardised, log-transformed sensor-by-feature matrix to identify a small number of components (linear combinations of the original temporal variables) that capture the dominant axes of variation in temporal NO_2 behaviour across sensors. Because the input matrix was already centred and scaled prior to PCA, the decomposition itself was performed without additional centring or scaling (centre = FALSE, scale. = FALSE in the prcomp call), avoiding redundant transformation of the data.

The proportion of variance explained (PVE) by each principal component, and the cumulative proportion of variance explained across the first k components, were used to guide the choice of how many components to retain for clustering. A scree plot of PVE by component indicated that the first six principal components captured a substantial share of the total variance in temporal NO_2 behaviour, with markedly diminishing returns from additional components; six components were therefore retained as the primary basis for clustering (npc_default = 6), a choice re-examined in Section 3.4 alongside an eight-component alternative. Loadings plots (variable-contribution and biplot visualisations) were used to interpret which temporal variables contributed most strongly to each retained component, for example, whether a given component was driven primarily by seasonal (monthly) structure or by diurnal/weekly structure, providing a qualitative check that the retained components corresponded to interpretable temporal patterns rather than arbitrary statistical artefacts.

3.5 Hierarchical Clustering and Selection of Cluster Number

Sensor scores on the retained principal components were used as the input to hierarchical clustering. Pairwise Euclidean distances between sensors were computed in the reduced principal-component space, and clusters were formed using Ward's minimum-variance linkage criterion (Ward's D2), which minimises the total within-cluster variance at each merge step and tends to produce compact, well-separated clusters.

Selecting the number of clusters, k, in unsupervised clustering is inherently a multi-criteria decision rather than one governed by a single definitive statistical test. We therefore used the NbClust package, which computes up to 30 internal validity indices (e.g., silhouette width, Calinski-Harabasz, Duda-Hart) across a candidate range of cluster numbers and reports the value of k favoured by a plurality of indices, as a quantitative starting point. This procedure was run separately using both eight and six retained principal components, evaluating candidate solutions from k = 2 to k = 8. In both cases, a plurality of indices favoured k = 3, though substantial minorities of indices favoured k = 7 or k = 8, reflecting the well-documented tendency of validity indices to disagree when cluster structure is only moderately distinct, which is common in continuous, environmentally driven data that need not partition into perfectly discrete groups.

Because the majority-rule recommendation from NbClust did not clearly dominate the alternatives, and because a coarser partition (k = 3) risked merging temporally and spatially distinct source signatures (for example, combining industrial/rail-proximate sensors with high-traffic-corridor sensors, which show different diurnal shapes despite comparable magnitude), we selected a final solution of k = 5 clusters. This choice was informed by inspection of the dendrogram structure, in which five clusters corresponded to a natural, well-separated cut point. It was validated using silhouette analysis, which confirmed that the five-cluster solution achieved meaningfully higher silhouette width than lower-k alternatives (0.42 vs. 0.2, respectively), though not indicative of extremely well-separated clusters in an absolute sense, which is consistent with the expectation that real-world pollutant time series vary along a continuum rather than falling into perfect categories.

Cluster assignments were obtained by cutting the dendrogram at the height corresponding to five groups (cutree), and each sensor was labelled with its resulting cluster identity for all subsequent analyses. Cluster robustness was additionally assessed qualitatively through pairwise scatterplots of sensors in principal-component space (e.g., PC1–PC2, PC2–PC3) overlaid with convex hulls by cluster, confirming that the five clusters occupied largely distinguishable regions of the reduced feature space rather than overlapping diffusely.

3.6 Buffer-Based Land Use and Infrastructure Characterisation

For each sensor, we characterised the surrounding built environment using the CMAP land-use inventory and TIGER/Line roadway and rail layers within three buffer radii (0.5, 1, and 2 km), generated in ArcGIS Pro using the Buffer tool. Land use area within each buffer was computed using the Tabulate Intersection tool, which calculates the area of each land-use category intersecting each buffer polygon; this required a manual reclassification step to address CMAP code 6000 (“Non-Parcel Area”), a residual category encompassing unclassified space that would have obscured meaningful land use signal. Non-parcel area was reclassified through spatial overlay with the roadway and hydrography reference layers described in Section 2.3, assigning portions of this category to roadway or water where a clear geometric overlap existed and retaining the remainder as genuinely unclassified. For each sensor x buffer combination, the top four land use categories by area share were retained (LANDUSE_1–LANDUSE_4), along with their percentage of total buffer area. The top four categories were retained because, in preliminary inspection, they accounted for the large majority of classified area in nearly every buffer; retaining additional lower-ranked categories added little further explanatory information while substantially increasing the dimensionality of the sensor metadata table.

Roadway characterisation used the Pairwise Intersect tool to clip TIGER/Line road segments to each buffer polygon, after which segment-level AADT values (by functional class) were averaged and segment lengths summed within each buffer, producing mean AADT and total road length by functional class (primary, secondary, local, and additional transportation/utility subcategories) for each sensor–buffer combination. Rail characterisation followed an analogous procedure, yielding total rail length and segment count within each buffer.

The resulting sensor metadata table was merged with cluster assignments by sensor name, enabling direct comparison of land-use composition, road traffic volume, and rail proximity across the five temporal clusters identified in Section 3.5. Because buffer radius affects which land uses and infrastructure fall within the sensor's local environment, all land use and infrastructure comparisons across clusters were conducted separately at each of the three buffer radii, allowing us to assess whether observed associations were consistent across spatial scales or were sensitive to the choice of buffer size.

3.7 Statistical Comparison of Clusters

To formally test whether NO_2 concentrations differed across the five clusters, we compared sensor-level mean NO_2 concentrations using the Kruskal-Wallis test, a non-parametric analogue to one-way ANOVA that does not assume normally distributed residuals, which is an appropriate test given the right-skewed nature of the underlying concentration data (Section 3.2). Where the Kruskal-Wallis test indicated a significant overall difference among clusters, pairwise comparisons were conducted using Dunn's test with Bonferroni correction for multiple comparisons ($\alpha = 0.05$), identifying which specific pairs of clusters differed significantly in mean NO_2 concentration. Land-use, AADT, and rail-length comparisons across clusters (Section 4) are reported descriptively (means and, where relevant, standard errors) given the exploratory nature of the built environment characterisation relative to the temporal clustering itself.

4 Results and Discussion

4.1 Five Clusters Emerge from the Sensor Network

Hierarchical clustering of the principal-component representation of each sensor's temporal NO_2 signature yielded five distinct groups, hereafter labelled C1 through C5, comprised of 89, 109, 21, 30, and 19 sensors, respectively (total $n = 268$). The unequal cluster sizes are themselves informative: the two largest clusters (C1 and C2, together accounting for roughly three-quarters of the network) capture the temporal behaviour typical of most of the city, while the three smaller clusters (C3, C4, C5) identify minority subsets of sensors whose NO_2 dynamics diverge from this baseline. This structure is consistent with an urban pollution landscape in which most locations experience a broadly similar background regime modulated by traffic and meteorology, while a smaller number of locations sit close enough to concentrated emission sources, or far enough from them, to display a qualitatively different signature.

A Kruskal-Wallis test confirmed that sensor-level mean NO_2 concentration differed significantly across the five clusters ($H = 206.5$, $p < 0.001$; Figure 1), and Dunn's post hoc test with Bonferroni correction resolved this omnibus difference into four statistically distinct concentration tiers. Ranked from highest to lowest mean NO_2, the clusters fall into groups {C3, C4}, {C1}, {C2}, and {C5}, meaning, C3 and C4 are not statistically distinguishable from one another despite C3's visibly higher median in Figure 1, while C1, C2, and C5 are each significantly distinct from every other cluster.

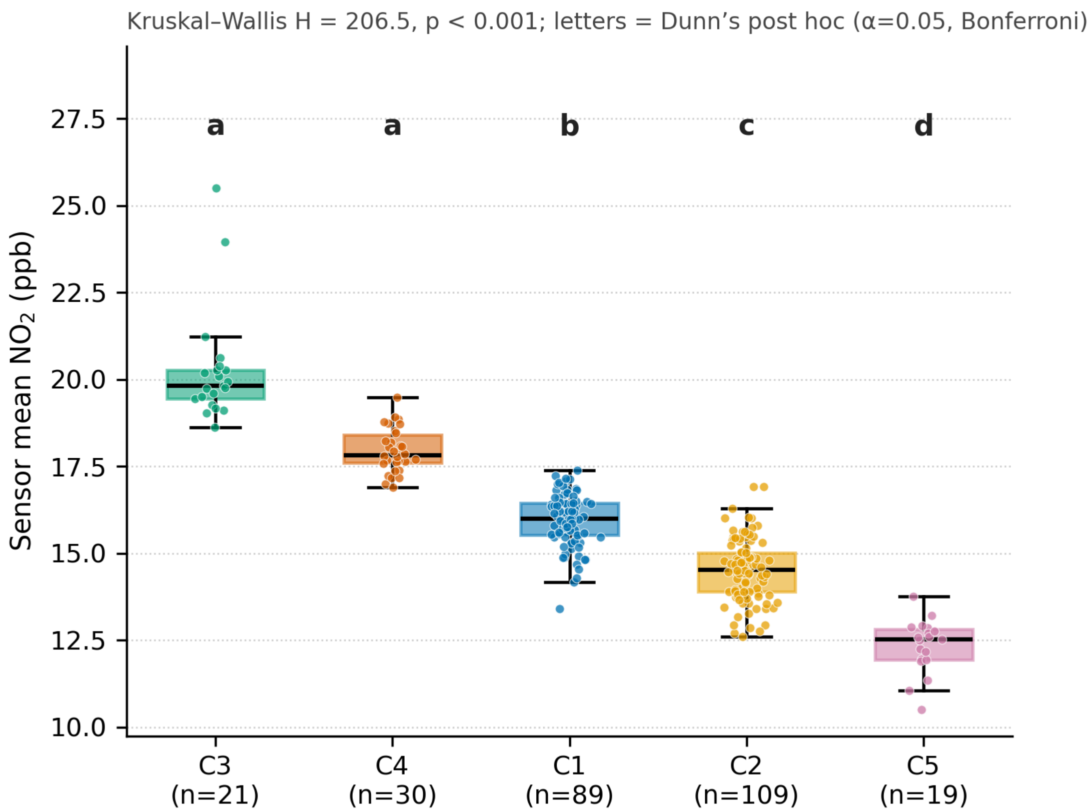

Figure 2. Distribution of sensor-level mean NO_2 concentration by cluster, ranked by mean concentration. Boxes show the interquartile range with the median indicated by the horizontal line; points are individual sensors. Letters above each cluster denote statistically distinct groups from Dunn's post hoc test with Bonferroni correction ($\alpha = 0.05$), following a significant Kruskal-Wallis test ($H = 206.5$, $p < 0.001$

The highest tier of clusters (C3/C4, sensor means clustering around 18-20 ppb) sit well above the lowest tier of cluster (C5, mean near 12-13 ppb), which is a gap of roughly 6-7 ppb, which is a meaningful fraction of typical urban background NO_2. This gap is large enough to matter for both compliance monitoring and environmental justice assessment, since a resident living near a C3 sensor may experience a materially different chronic exposure regime than one near a C5 sensor, independent of any city-wide average concentration reported for regulatory purposes. Even so, this gap should be read against applicable health-based benchmarks: the nine-month mean NO_2 concentration at every sensor in the network, including those in the highest-concentration cluster, remains well below the U.S. EPA's annual NO_2 NAAQS of 53 ppb, and it is only the WHO's considerably more stringent 2021 annual guideline of approximately 5 ppb (10 µg/m³) that most of the network, including the lowest-concentration cluster, exceeds (World Health Organization, 2021).

4.2 Distinct Diurnal, Weekly, and Seasonal Signatures Distinguish the Clusters

The relative NO2 concentrations of the clusters remain consistent across all timeframes examined, with cluster C3 having the highest NO2 during all hours, weekdays, and months, and C5 having the lowest NO2.

The diurnal profile of mean NO_2 by cluster (Figure 2) shows that all five clusters share a common qualitative shape -- a pronounced early-morning peak (approximately 05:00-06:00), a midday trough, and a broader evening peak (approximately 18:00-20:00), consistent with a signature dominated by traffic combined with the diurnal cycle of the atmospheric boundary layer, which tends to trap pollutants near the surface overnight and during the early-morning and again as the layer collapses in the evening. What differs across clusters is not the timing of these features but their relative amplitude and persistence: C3 and C4 show the sharpest, most pronounced morning peaks and maintain elevated concentrations throughout the midday relative to C1, C2, and C5, whose midday concentrations drop closer to a shared floor. This pattern is characteristic of locations with a more continuous, less strictly commute-timed emission source superimposed on the traffic-driven cycle common to the whole network, such as intermodal freight operations or industrial activity that operates across daytime hours rather than concentrating narrowly at rush hour.

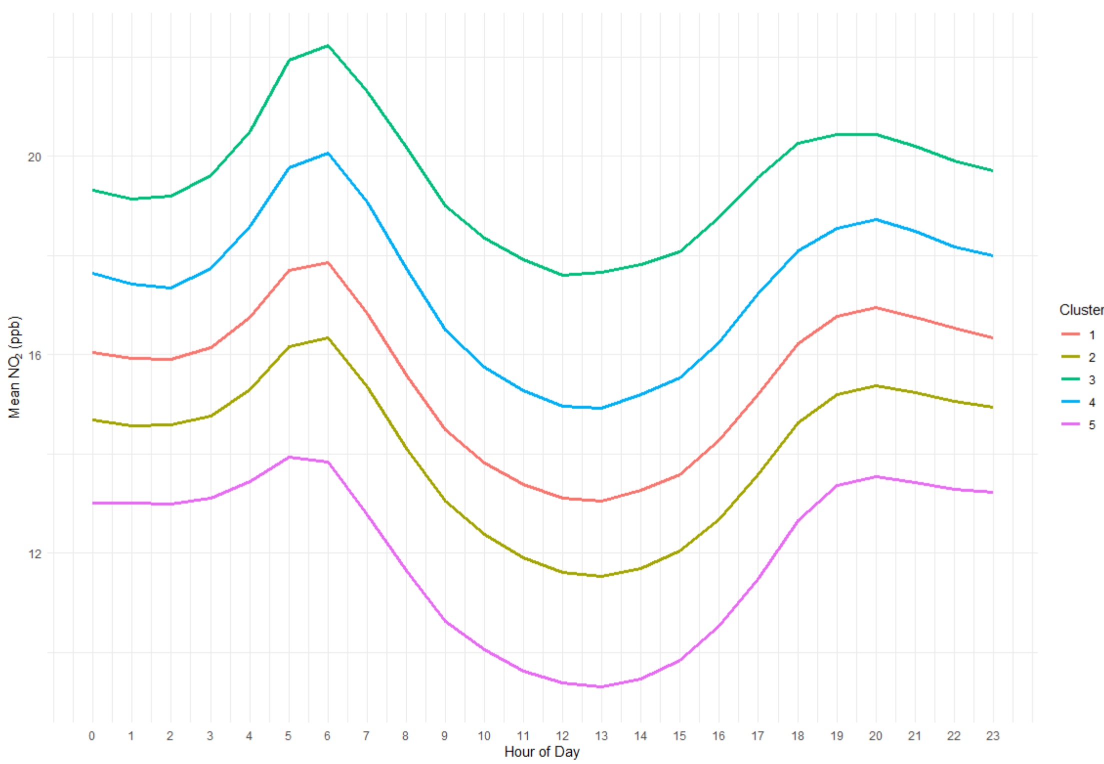

Figure 3. Mean NO_2 concentration by hour of day for each cluster

The day of week comparison (Figure 3) reinforces this interpretation. Concentrations decline from weekday to weekend across all five clusters, but the magnitude of the weekday-to-weekend drop is not uniform. C3 shows the steepest weekend decline in absolute terms, falling from roughly 20-21 ppb on weekdays to about 17-17.5 ppb on Saturday/Sunday, while C5 shows a comparatively modest and proportionally similar decline from an already low baseline. A steep weekday-to-weekend drop in the highest-concentration cluster is consistent with a source profile influenced by sources that decrease their activity substantially on weekends, rather than a source that operates on a fixed schedule independent of the workweek.

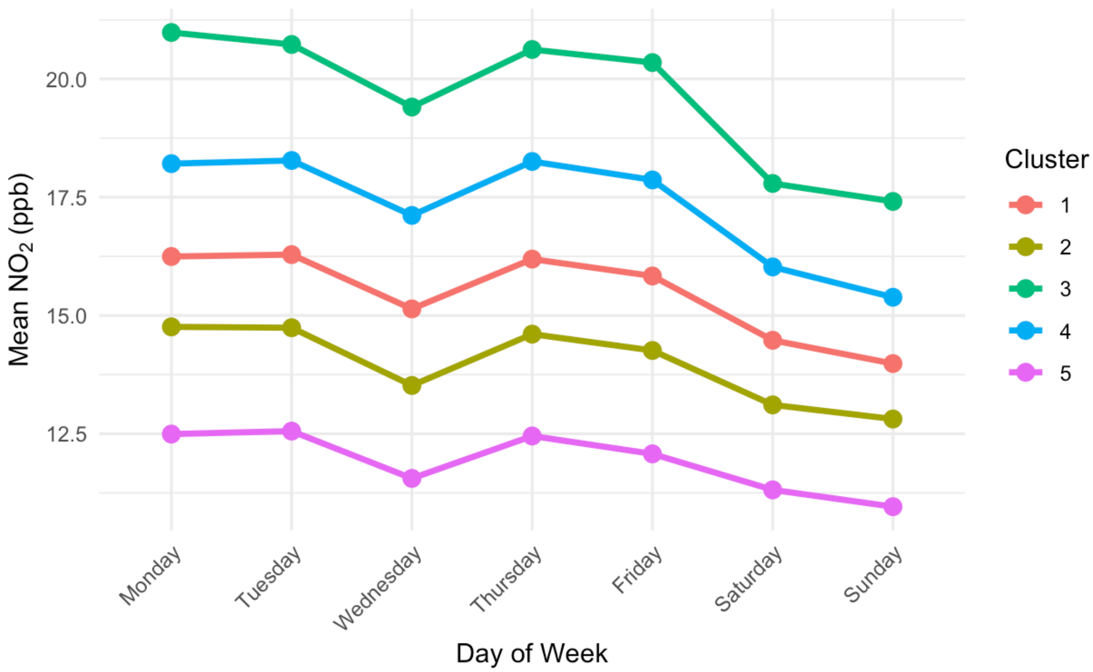

Figure 4. Mean NO_2 concentration by day of week for each cluster

The winter months carried the lowest sensor data completeness of the study period (Figure 2), so the following seasonal comparisons, particularly the December-February peak, are based on comparatively fewer complete sensor-months than the fall and spring estimates and should be interpreted with this reduced sample size in mind. The monthly time series (Figure 4) adds a seasonal dimension to this picture. All clusters show a pronounced cool season peak rising from September into a maximum around December-February before declining through spring, which is consistent with well-established seasonal NO_2 chemistry. NO_2 has a shorter atmospheric lifetime in summer, when higher temperatures and stronger photochemistry accelerate its conversion to ozone and other oxidation products, and a longer lifetime in winter, when a shallower, more stable boundary layer limits vertical dispersion and slower photochemistry allows NO_2 to accumulate closer to its emission sources. What is notable here is that this seasonal amplitude is not uniform across clusters: C3 rises from roughly 18-19 ppb in early fall to a winter peak near 26 ppb before falling back to roughly 18-19 ppb by spring (a swing of nearly 8 ppb) while C5 shows a comparatively smaller swing of only 2-3 ppb around a much lower baseline.

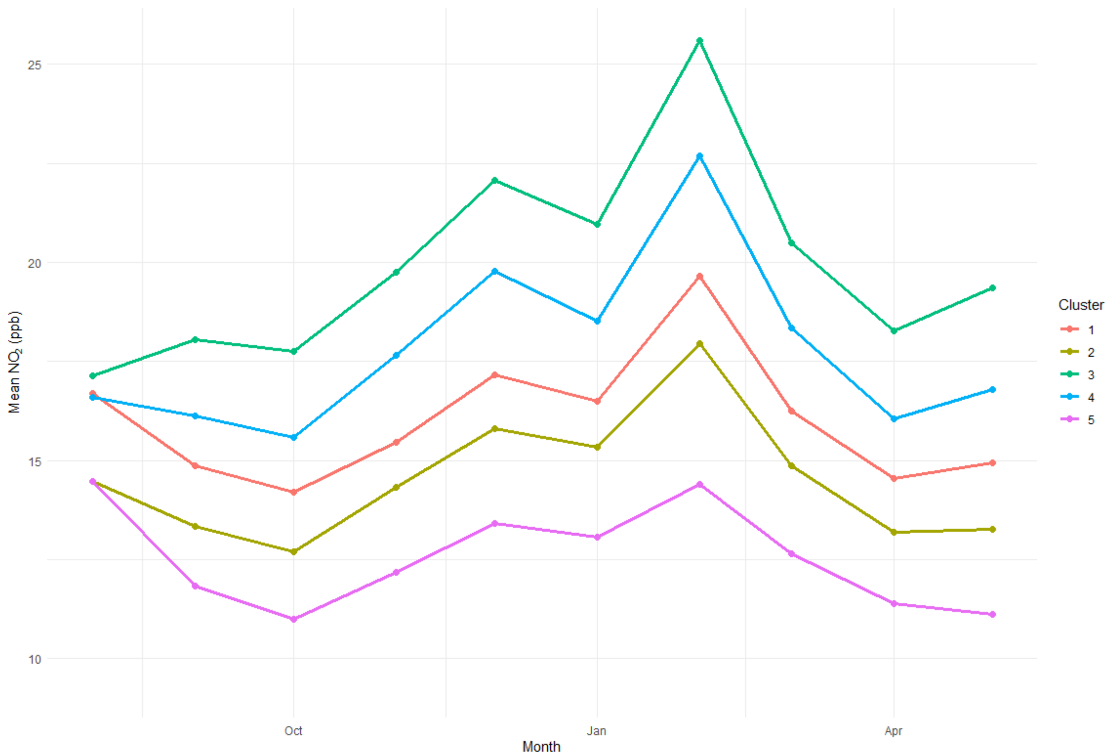

Figure 5. Mean monthly NO_2 concentration by cluster, September 2025 through May 2026

The internal consistency in cluster concentrations across independent temporal dimensions (hour, weekday, month) strengthens confidence that the five-cluster solution selected in Section 3.4, rather than the three-cluster solution, captures meaningful distinctions that a coarser partition would have obscured.

4.3 Land Use and Infrastructure Explain the Cluster NO2 Patterns

The central objective of this analysis was to determine whether the clusters derived from the NO_2 measurements, without incorporating any spatial information during clustering, corresponded to meaningful differences in the surrounding built environment. The results indicate that the clusters are not randomly distributed across the urban landscape but instead align with distinct patterns of land use and transportation infrastructure. In particular, the highest-concentration clusters are characterised by greater industrial activity and freight-related transportation infrastructure, suggesting that differences in the type and context of nearby emission sources contribute to the observed spatial variability in NO_2 concentrations.

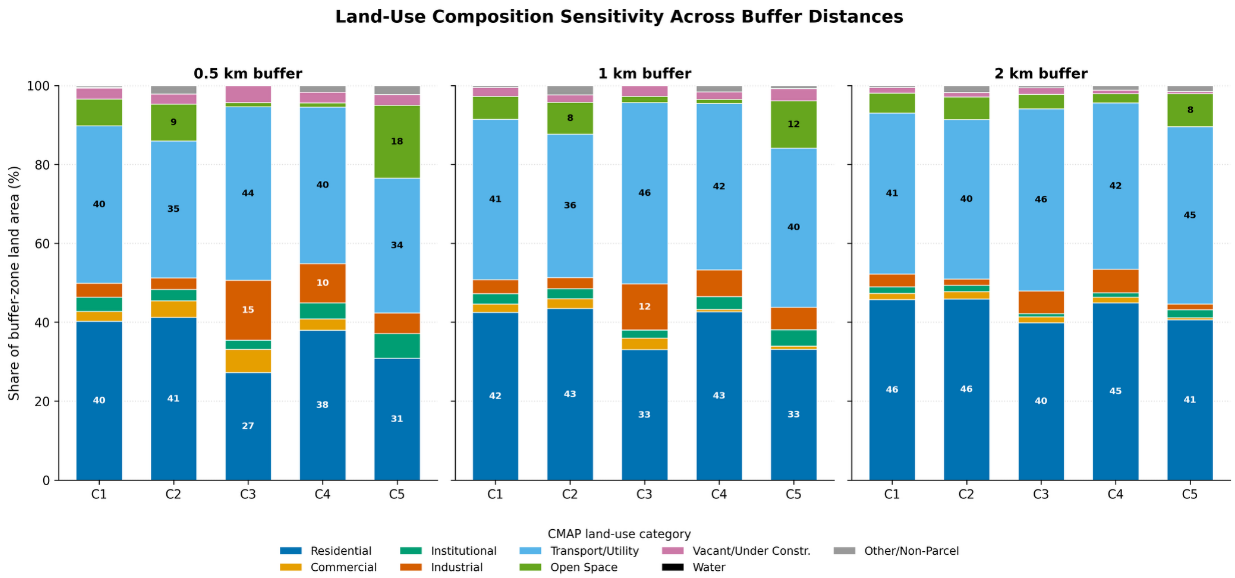

Figure 6. CMAP land-use composition by cluster and buffer distance (0.5, 1, and 2 km). Values are percentages of total buffer-zone land area.

The land-use composition surrounding each cluster varied depending on the spatial scale used to characterise the sensor environment. At the smallest buffer distance (0.5 km), the differences between clusters were most apparent, with C3 and C4 exhibiting greater industrial land-use contributions compared with the remaining clusters. Industrial land use accounted for approximately 15% of the area surrounding C3 sensors and 10% surrounding C4 sensors, whereas C1, C2, and C5 contained minimal industrial land use within the same buffer. This pattern is consistent with the concentration-based clustering results, where C3 and C4 represented the highest NO_2 concentration tier across the network.

However, the magnitude of these differences decreased as the buffer distance increased. At the 1 km and 2 km scales, the surrounding landscape increasingly reflected the broader urban composition of Chicago, with residential and transportation-related land uses becoming dominant across clusters. This suggests that buffer size influences the ability to identify localised source characteristics, with smaller buffers better preserving the nearby features that distinguish high- and low-concentration environments. Larger buffers, while useful for describing general neighbourhood context, increasingly average across the heterogeneous mixture of land uses surrounding urban sensors.

The relationship between industrial land use and elevated NO_2 concentrations should be interpreted within the broader context of the urban environment. Although industrial land use represents only a relatively small fraction of total surrounding land area compared with residential and transportation categories, its consistent association with the highest concentration clusters suggests that localised industrial activity may provide important context for understanding areas with elevated NO_2. Importantly, industrial land use alone does not explain the observed concentration patterns; rather, it occurs alongside a combination of transportation and freight-related infrastructure that further distinguishes the highest concentration environments.

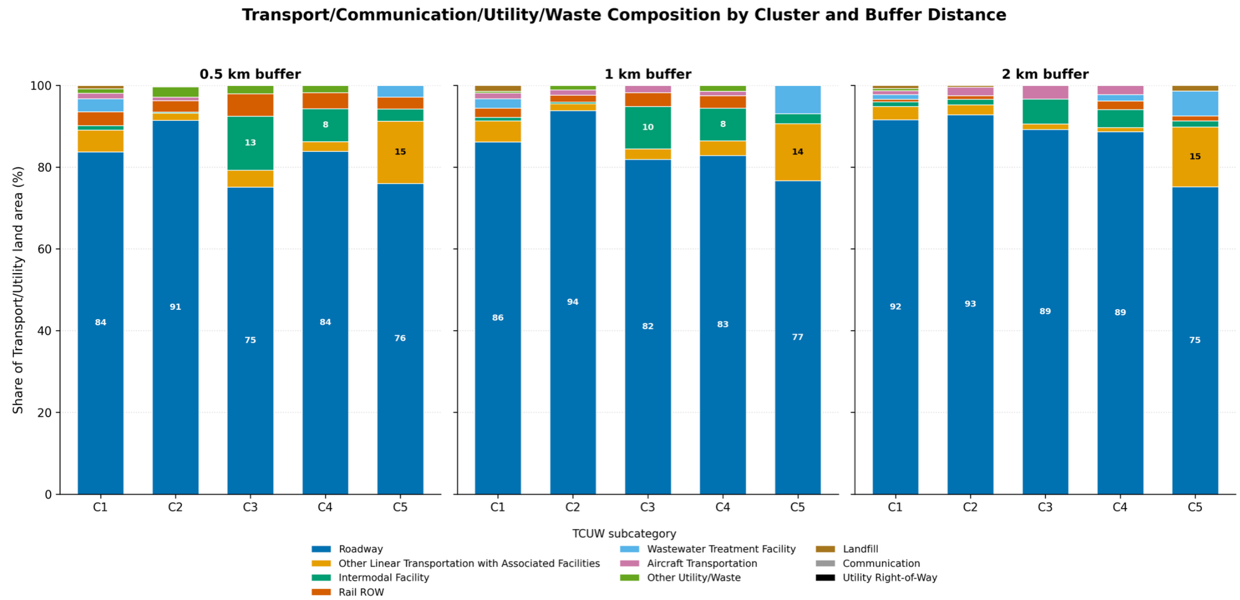

Figure 7. Composition of total transportation/communication/utility/waste (TCUW) land use, including roadway, by cluster and buffer distance

Transportation-related land use was further examined by separating the transportation, communication, utility, and waste (TCUW) category into its individual components. Roadway land use represented the dominant component across all clusters and buffer distances, accounting for approximately three-quarters to more than ninety percent of the total TCUW category. This widespread presence reflects the extensive roadway network throughout Chicago and highlights the importance of transportation emissions as a ubiquitous feature of the urban environment.

However, because roadway land use was consistently dominant across all clusters, it provided limited ability to explain why some locations exhibited substantially higher NO_2 concentrations than others. The greater differences between clusters were contained within the smaller fraction of transportation-related land use associated with non-roadway infrastructure, including rail corridors, intermodal facilities, and other transportation-related land uses. These components represent a more spatially variable subset of the transportation environment and therefore provide greater insight into the characteristics distinguishing higher- and lower-concentration clusters.

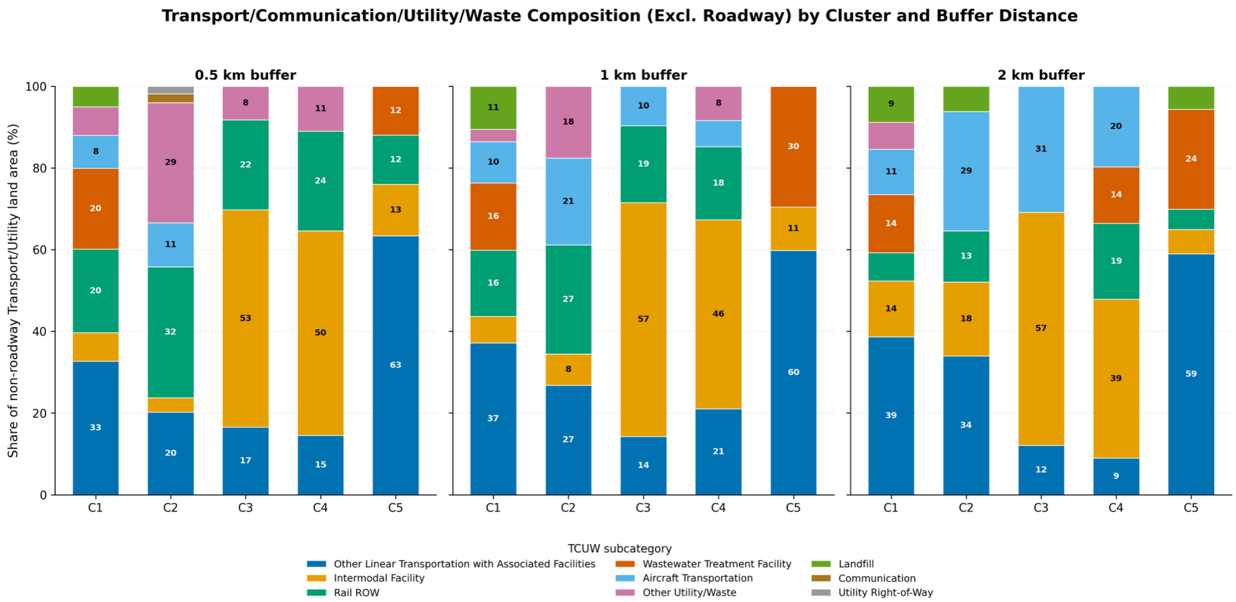

Figure 8. Composition of non-roadway transportation, communication, utility, and waste (TCUW) land use by cluster and buffer distance (0.5, 1, and 2 km). Values are percentages of non-roadway TCUW land area

Because transportation-related land use represented the largest overall land-use category across clusters, this category was further examined by separating roadway infrastructure from the remaining transportation-related land uses. Roadways accounted for approximately three-quarters to more than ninety percent of the total transportation, communication, utility, and waste (TCUW) category across buffer distances. This widespread dominance reflects the extensive roadway network throughout Chicago but provides limited ability to distinguish between clusters because roadway infrastructure is common throughout the study area.

When roadway land use was removed, clearer differences emerged between clusters. The remaining transportation-related land-use composition revealed substantially different infrastructure profiles, particularly between the highest and lowest concentration groups. C3, the highest concentration cluster, was dominated by rail right-of-way and intermodal facilities, which together accounted for approximately 85% of the remaining non-roadway transportation land use within the 0.5 km buffer. C4 exhibited a similar pattern, although less pronounced, with greater representation of freight-related transportation infrastructure compared with the lower concentration clusters.

Intermodal facilities represent a particularly important distinction because they function as locations where goods are transferred between transportation modes, commonly involving interactions between freight trains, heavy-duty trucks, and cargo-handling equipment. These activities represent a combination of transportation sources that differs from simple proximity to a roadway or rail corridor alone. The elevated NO_2 concentrations observed in C3 and C4 are therefore consistent with locations where multiple emission sources associated with freight movement occur within close proximity.

The comparison between C2 and C3 further highlights this distinction. Although C2 contained a substantial proportion of rail infrastructure, it did not exhibit the same elevated NO_2 concentrations as C3. This suggests that rail proximity alone may not be sufficient to explain concentration differences across the network. Instead, the combination of rail infrastructure with active freight-handling activities, particularly intermodal operations, appears to provide additional information regarding the source environments associated with the highest measured NO_2 concentrations.

4.4 Sensitivity to Buffer Radius

Because the choice of buffer radius is a methodological decision rather than a physical constant, it is important to question whether the associations identified above depend on which buffer was analysed. Figure 8 provides this sensitivity check by presenting land-use composition at 0.5, 1, and 2 km simultaneously. Two patterns emerge. First, the broad rank-ordering of clusters by industrial and open-space land-use share is preserved across all three radii: C3 remains the cluster with the most non-residential, non-transport land-use signal (industrial, at smaller buffers) at every scale, and C5 retains an elevated open-space share at every scale, though the magnitude of these differences compresses somewhat as buffer radius increases, since a larger buffer averages in more of the surrounding heterogeneous mix and dilutes any sharply localised signal. Second, the residential land-use share converges toward a narrower range across clusters as buffer radius increases (27-41% at 0.5 km versus 40-46% at 2 km), reflecting the simple geometric fact that Chicago is a residentially dominated city at coarse spatial scales, and that fine-scale industrial or transportation signal is progressively „averaged out“ as the buffer size increases to encompass more of the surrounding, more homogeneous residential mix.

This sensitivity pattern has a clear practical implication for future work using this methodology: analysis seeking to characterise the proximate emission source most plausibly responsible for a sensor's elevated readings should favour smaller buffers (0.5-1 km), where source-specific land-use signal is strongest and least diluted, while larger buffers (2 km) are better suited to characterising the general neighbourhood context in which a sensor sits. The fact that our core qualitative findings of C3's rail/intermodal association and C5's open-space/limited-access-roadway association hold across all three buffer scales, even as their magnitude attenuates, indicates these are not artefacts of an arbitrarily chosen buffer size but reflect spatially persistent features of the built environment near these sensors.

Figure 9, which presents the same TCUW composition including roadway land use, reinforces this robustness assessment from a different angle. Roadway land use dominates the transport/utility category everywhere (75-94% of TCUW area across clusters and buffers), which is unsurprising in a dense urban street grid, but the residual 6-25% of TCUW land not devoted to ordinary roadway is where the cluster-differentiating signal concentrates (Figure 6), and this residual, non-roadway signature is what most cleanly separates C3's rail/intermodal profile from C5's limited-access-roadway profile. This is a useful methodological observation for future dense-sensor-network studies: when roadway land use is nearly ubiquitous across a study area, isolating the non-roadway „residual" transportation and industrial land use may be a more discriminating feature for source attribution.

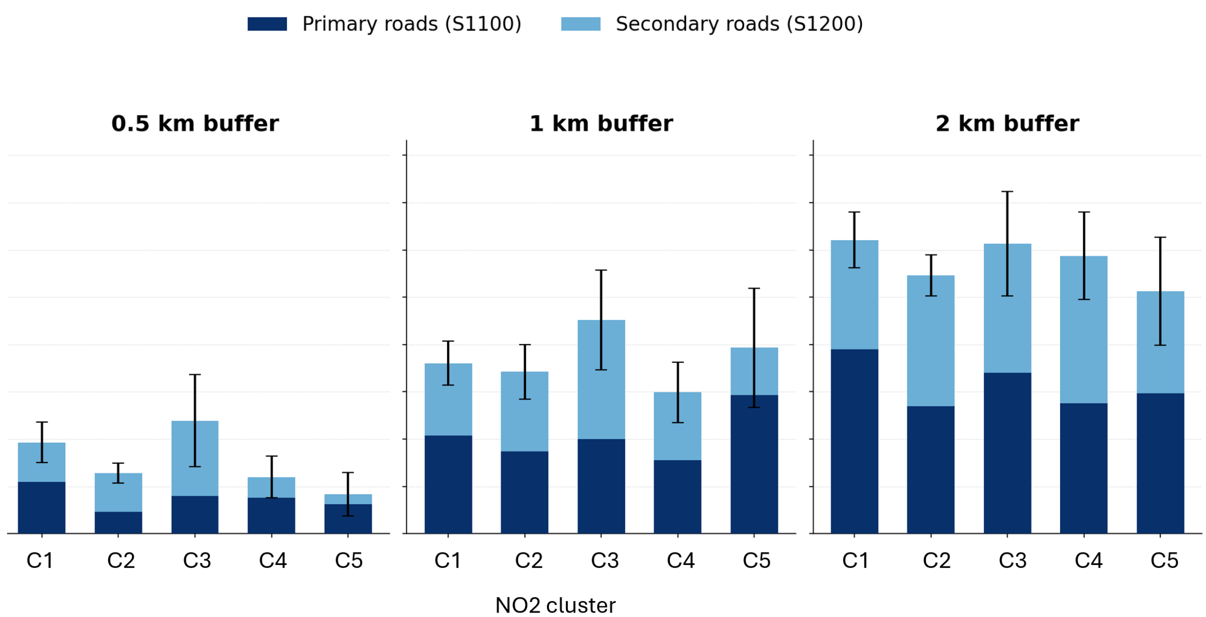

Figure 9. Mean annual average daily traffic (AADT) on primary (interstate/expressway) and secondary (arterial/state highway) roads within each buffer of each sensor, by cluster

The relationship between traffic volume and NO_2 concentrations was further examined using average annual daily traffic (AADT). If traffic volume alone explained the observed concentration patterns, the highest concentration clusters would be expected to consistently exhibit the greatest nearby traffic volumes. However, this relationship was not observed. C5, which consistently represented the lowest NO_2 concentration cluster, exhibited comparable or greater primary roadway traffic volumes than C3, particularly at larger buffer distances.

This finding does not indicate that roadway emissions are unimportant. Roadway traffic remains a major contributor to urban NO_x emissions and influences NO_2 concentrations throughout the network. Rather, the results demonstrate that traffic volume alone is an incomplete indicator of localised NO_2 variability. Similar traffic volumes may correspond to substantially different concentration environments depending on roadway characteristics, surrounding land use, proximity to freight activity, and the presence of additional co-located emission sources.

The contrasting profiles of C3 and C5 illustrate this distinction. C5 contained substantial roadway influence but lacked the freight-related and industrial characteristics associated with the elevated concentration clusters. In contrast, C3 combined roadway infrastructure with rail, intermodal, and industrial activity, creating a more complex transportation environment with multiple potential NO_2 sources. These results indicate that transportation exposure characterisation should extend beyond vehicle counts alone and incorporate the broader context in which transportation activity occurs.

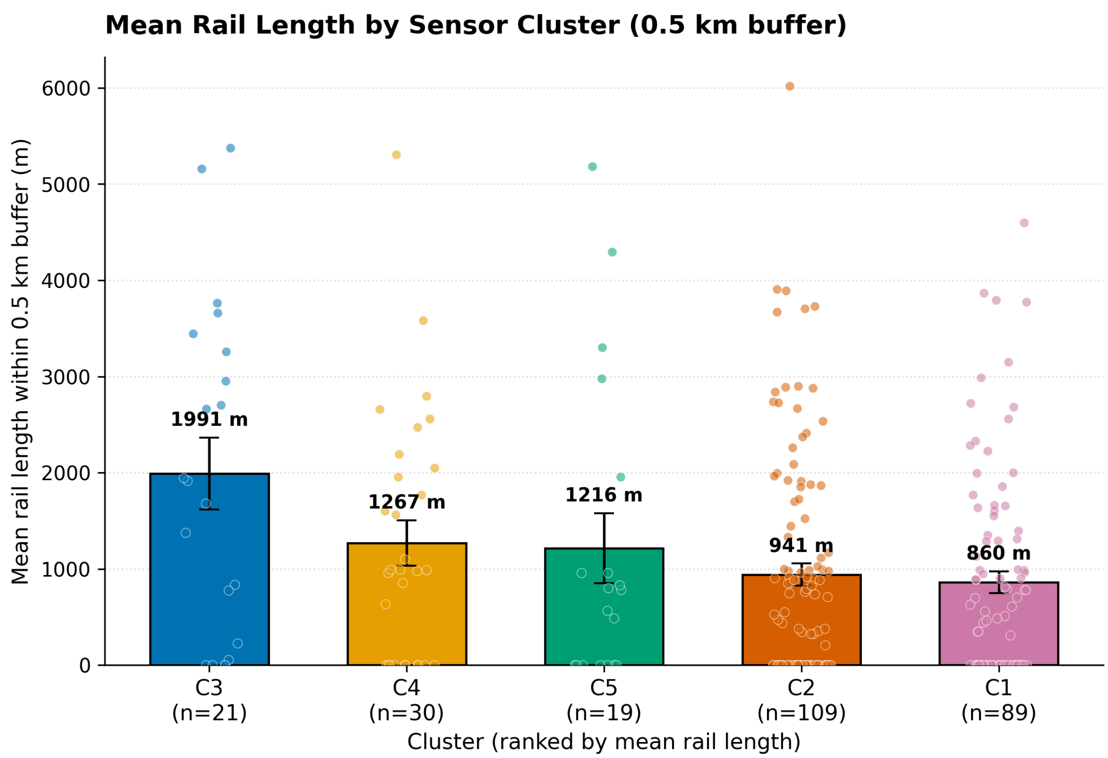

Figure 10. Mean rail infrastructure length within 0.5 km of sensor location, by cluster (ranked by mean rail length). Points show individual sensors; error bars represent standard error of the mean

The importance of freight-related infrastructure was further evaluated by examining rail infrastructure surrounding each cluster. C3 contained the greatest average rail length within 0.5 km of sensors, approaching 2 km, which was substantially greater than the other clusters. C4 also exhibited elevated rail infrastructure compared with the lower concentration clusters, providing additional evidence that the highest NO_2 environments are associated with greater freight-related infrastructure.

However, like the comparison between C2 and C3, rail length alone does not fully explain the observed concentration differences. C2 contained meaningful rail infrastructure but did not exhibit the elevated NO_2 concentrations observed in C3. Rail length provides information about the physical presence of infrastructure but does not capture the intensity, type, or timing of activities occurring within those locations.

Taken together, these results suggest that elevated NO_2 concentrations are associated not simply with transportation infrastructure in general, but with specific combinations of transportation activity and surrounding land use. Areas characterised by the co-occurrence of rail infrastructure, intermodal facilities, industrial land use, and roadway access represent distinct source environments that are more strongly associated with the highest concentration clusters observed across the network.

The results demonstrate that temporal patterns in NO_2 concentrations identified through unsupervised clustering correspond to meaningful differences in the surrounding urban environment. Importantly, the clustering approach was performed independently of land-use and infrastructure information, meaning that the observed relationships represent an external validation that sensors with similar NO_2 behavior are also located in areas with similar physical characteristics.

Across the network, transportation-related land use was the dominant surrounding category, reflecting the highly urbanized nature of Chicago. However, the results demonstrate that presence of transportation alone does not explain the observed concentration differences between clusters. Roadways were widespread across all clusters, and traffic volume alone did not consistently correspond with higher measured NO_2 concentrations. Instead, the strongest differentiation occurred in the specific types of transportation infrastructure present, particularly the combination of rail, intermodal facilities, industrial land use, and freight activity surrounding the highest concentration clusters.

This distinction is important because roadway emissions remain a major contributor to urban NO_2 concentrations, but their influence is shaped by the surrounding source environment. Locations with similar levels of roadway traffic can experience different concentration regimes depending on whether traffic occurs alongside additional emission sources such as freight-handling operations or industrial activity. The results therefore suggest that assessments of near-roadway NO_2 variability should consider not only traffic intensity but also the broader transportation system in which roadways are embedded.

The environmental implications of these differences are also important. The approximately 6–7 ppb separation between the highest and lowest concentration tiers represents a meaningful difference in long-term neighbourhood-scale NO_2 exposure. While this analysis does not directly evaluate population exposure or demographic disparities, it demonstrates that some communities are located in environments where residential land use overlaps with industrial, transportation, and freight-related activities. Combining dense sensor observations with detailed spatial information therefore provides a useful framework for identifying locations where additional exposure assessment or mitigation efforts may be warranted.

A further methodological finding from this analysis is the importance of spatial scale when interpreting sensor environments. The strongest separation between clusters was observed at the 0.5 km buffer, where localised industrial and transportation characteristics were most apparent. As buffer size increased, these differences became increasingly diluted as the analysis captured a broader urban mixture. This suggests that smaller spatial scales may be particularly useful for identifying nearby emission source characteristics, while larger buffers may better represent general neighbourhood context.

Overall, this study demonstrates the value of integrating dense sensor networks with spatial datasets to investigate urban air-quality variability. Sensor measurements provide high-resolution information about where concentrations differ, while land-use and infrastructure data provide insight into the characteristics associated with those differences. Although additional analyses incorporating meteorology, emissions inventories, and more detailed indicators of transportation activity are needed to move from association toward source attribution, the results demonstrate that combining these approaches can reveal spatial patterns in NO_2 concentrations that would not be apparent from either dataset alone.

4.5 Limitations and Comparison with Expectations

Several limitations should be considered when interpreting these findings. First, low-cost sensors introduce greater measurement uncertainty compared with reference-grade monitoring instruments. Although the observed spatial patterns are internally consistent across multiple temporal dimensions and align with independent land-use characteristics, further evaluation against regulatory monitoring networks and quality-control records is needed to fully characterise measurement uncertainty.

Second, the land-use analysis relies on a static, buffer-based representation of the surrounding environment. This approach assumes that proximity provides a useful approximation of source influence but does not account for meteorological factors such as wind direction, atmospheric stability, or time-varying source activity. The persistence of strong associations between NO_2 clusters and surrounding land-use characteristics despite these limitations suggests that static proximity provides a useful representation of source environments at the multi-month averaging scale. However, incorporating meteorological information and wind-resolved source influence would likely improve understanding of the relationship between emissions and observed concentrations.

Third, although the five-cluster solution provided improved separation compared with coarser clustering approaches, the silhouette width of 0.42 indicates moderate rather than complete separation between groups. Some sensors located near cluster boundaries in principal-component space may therefore represent intermediate conditions rather than distinct source environments.

Finally, the completeness filtering applied during preprocessing may have influenced the retained sensor population. Removing sensors and variables below the 50% completeness threshold may disproportionately exclude newly deployed or intermittently reporting sensors, potentially biasing the analysis toward more established monitoring locations. Future work should evaluate whether excluded sensors differ systematically from retained sensors in terms of geographic distribution or surrounding land-use characteristics.

These findings are broadly consistent with the expectations motivating this study (Section 1): that traffic-adjacent and freight/industrial-adjacent locations would show both elevated concentrations and distinctive temporal signatures relative to more residential, lower-traffic locations.

5 Conclusion

This study demonstrates that dense, low-cost NO_2 sensor networks can be used to identify distinct urban emission environments by combining temporal concentration patterns with independent spatial characterisation of the surrounding built environment. Clustering the Open Air Chicago sensor network using NO_2 temporal signatures alone identified five groups with significant differences in concentration levels, and these clusters aligned with meaningful differences in land use and transportation infrastructure.

The highest concentration clusters were distinguished not simply by roadway proximity or traffic volume, but by the presence of freight-related infrastructure, particularly rail and intermodal facilities, combined with greater industrial land-use representation. In contrast, locations with substantial roadway traffic but limited freight infrastructure did not consistently exhibit elevated NO_2 concentrations. These findings indicate that roadway emissions remain an important component of the urban NO_2 environment, but that traffic volume alone does not capture the full complexity of localised concentration patterns. The surrounding combination of transportation type, freight activity, and industrial context appears to be critical for understanding why some areas experience persistently higher concentrations than others.

These findings have important implications for air-quality management strategies. Approaches focused solely on reducing overall traffic volumes may not fully address the locations experiencing the highest measured NO_2 concentrations if those locations are also influenced by freight-related activities such as intermodal operations, rail activity, and diesel drayage. Strategies that target these source environments more directly may therefore provide a more effective pathway for reducing localised NO_2 exposure.

The approximately 6-7 ppb difference between the highest and lowest concentration tiers also highlights the importance of considering neighbourhood-scale variability in urban air quality assessment. Although these concentrations remain below the current U.S. EPA annual NO_2 standard, persistent differences of this magnitude represent meaningful differences in chronic exposure conditions between communities. This has particular relevance from an environmental justice perspective, as combining dense monitoring networks with land-use information can help identify locations where residential communities overlap with transportation and industrial emission sources and where additional investigation may be warranted.

More broadly, this study demonstrates the value of integrating high-resolution sensor observations with spatial datasets. Sensor networks provide detailed information on where and when concentrations differ, while land-use and infrastructure data provide context for interpreting the potential sources associated with those differences. Future work should incorporate meteorological information, facility-level emissions data, and more detailed measures of transportation activity to further evaluate source attribution and improve understanding of localised urban NO_2 variability.

6 References

Camilleri, Sara F., et al. "All-Cause NO2-Attributable Mortality Burden and Associated Racial and Ethnic Disparities in the United States." Environmental Science & Technology Letters, vol. 10, no. 12, 2023, pp. 1159-1164.

City of Chicago. "Open Air Chicago." Chicago Department of Public Health, 2026, www.chicago.gov/city/en/depts/cdph/supp_info/Environment/open-air-chicago.html.

de Foy, B. "City-Level Variations in NOx Emissions Derived from Hourly Monitoring Data in Chicago." 2018.

Dickens, A. F., et al. "Overview of the 2023 AGES+ Chicago Field Experiment."

Goldberg, D. L., et al. "Evaluating the Spatial Patterns of U.S. Urban NOx Emissions Using TROPOMI NO2." 2024.

Kerr, G. H., et al. "Air Pollution Impacts from Warehousing in the United States Uncovered with Satellite Data." 2024.

Lake Michigan Air Directors Consortium (LADCO). "Ozone Conceptual Model Report." www.ladco.org/wp-content/uploads/Projects/Ozone/Ozone-conceptual-model-report-FINAL-Feb-2023.pdf.

Lake Michigan Air Directors Consortium (LADCO). "Ozone Precursor Emissions Inventory Report." 2026, www.ladco.org/reports/ladco-2022-emissions-report/.

Lang, A., et al. "Intercomparison of Modeled Urban-Scale Vehicle NOx and PM2.5 Emissions -- Implications for Equity Assessments." 2025.

Miech, Jason A., et al. "Calibration of Low-Cost NO2 Sensors through Environmental Factor Correction." Toxics, vol. 9, no. 11, 2021, p. 281.

Shusterman, A. A., et al. "Observing Local CO2 Sources Using Low-Cost, Near-Surface Urban Monitors." 2018.

World Health Organization. WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. World Health Organization, 2021.