Uncovering Calendar-Based Structures Through Clustering of Daily Traffic Volume Profiles*

Wai Yu Amanda Ng

1 Introduction

Understanding traffic flow patterns is crucial for effective transportation management, urban planning, and congestion mitigation (Boyce 2012). To extract meaningful insights from dense traffic volume data which is usually continuously collected across multiple locations, effective dimensionality reduction techniques are required. This study aims to uncover the recurring temporal patterns and cross-location similarities in daily traffic volume by identifying clusters of days throughout the year that exhibit comparable traffic flow behavior across multiple locations. The study is conducted on a dataset capturing traffic volume across 26 locations for over a year. We first apply Principal Component Analysis (PCA) separately to each location to summarize dominant features in the location-specific daily traffic curves. We then combine these reduced representations across all locations and apply Uniform Manifold Approximation and Projection (UMAP), a nonlinear dimension reduction technique, on them to visualize and identify potential groupings of days with similar daily traffic volume characteristics. The PCA analysis reveals that location-specific daily traffic volume profiles exhibit some consistent structures across days, and UMAP results show that there are clusters of days corresponding to different school calendars, holiday periods and seasonal routines that show similar daily traffic volume patterns across locations. These findings provides insights to forcast fluctuations in traffic usage, hence aid tailoring transportation planning and infrastructure maintenance interventions.

*Editorial refinement of this manuscript was assisted by Microsoft Copilot (Microsoft 2025), which was used exclusively for polishing the writing style; all analyses, interpretations, and conclusions remain the sole responsibility of the author.

2 Data Description

2.1 Data Overview

The dataset used in this study is a synthetic dataset constructed based on a real-world traffic dataset analyzed in "Network-level traffic flow prediction: Functional time series vs. functional neural network approach" (Ma, Yao, and Zhou 2024). The original real data consists of two years of traffic volumes that were collected from the inductive loop detectors at 51 locations on Highway 401 in the City of Toronto, and it is aggregated to 288 data points per day at five-minute intervals. Due to confidentiality restrictions, this study uses the synthetic dataset, which replicated the pattern and covariance structure of the original data, hence preserving the real data's core characteristics. The synthetic dataset records daily traffic volumes at 26 locations over 384 days, measured every 5 minutes across a 24-hour period. Each location's traffic volume is represented by a matrix of size 384×288, where rows correspond to days and columns to time points. The (d, t) entry denotes the number of vehicles passing the location at day d and time index t (vehicles per 5 minutes).

2.2 Data cleaning

We confirmed that there is no missing entries in the dataset. To reduce initial dimensionality, the original 5-minute traffic counts (288 time points per day) were aggregated to hourly counts, resulting in 24 time points per day. This smooths short-term noisy fluctuations while preserving the broader key traffic patterns. Furthermore, only the first 365 days data was retained to focus our analysis within a one-year period. Each location's cleaned dataset therefore contains 365×24 (day $\times$ hour) observations, each capturing the traffic volumn within a specific hour on a specific day.

2.3 Data summary and visualizations

Table 1: Summary Statistics of Hourly Traffic Volume by 3-Hour Time Windows

Time Window	Mean	Standard Deviation
Late Night (00–03)	701.77	171.77
Early Morning (03–06)	1216.81	273.33
Morning Rush (06–09)	5032.83	1133.31
Late Morning (09–12)	4474.00	1039.81
Midday (12–15)	4612.63	1070.27
Afternoon Rush (15–18)	4545.31	1152.79
Evening (18–21)	3702.24	929.83
Night (21–24)	2279.69	571.30

In Table 1, for each time window, hourly traffic volumes were averaged across the corresponding hours, then averaged across all 365 days and all locations. We observe pronounced peaks during the morning and afternoon rush periods and lower traffic activity during late night hours, which is consistent with general public's daily traffic use pattern. Since some groups of time windows exhibit similar traffic volume statistics, this indicates potential redundancy in the original hourly data and dimensionality reduction may simplify the data structure effectively. As revealed from the daytime windows' higher standard deviations, the traffic behavior varies considerably across locations and days during the daytime. This may be driven by seasonal weather changes, differences between workdays and holidays, or location-related factors such as residential versus highway regions. Therefore, location-specific and day-specific variations should be taken into account when dimensionality reduction is conducted to ensure that these subtle patterns are not lost.

Figure 1 reveals broadly similar daily traffic volume patterns across locations, with elevated volumes during the morning and afternoon hours, and lower volumes during the early morning and late-night periods, aligning with the aggregated behavior shown in Table 1. However, the magnitude and shape of these peaks vary across locations. Most locations show a pronounced morning peak around 8AM, followed by a modest dip near noon. Some, but not all locations, show a secondary rise in the late afternoon, typically around 4PM, before tapering off into the evening. These variations suggest that while locations share a common

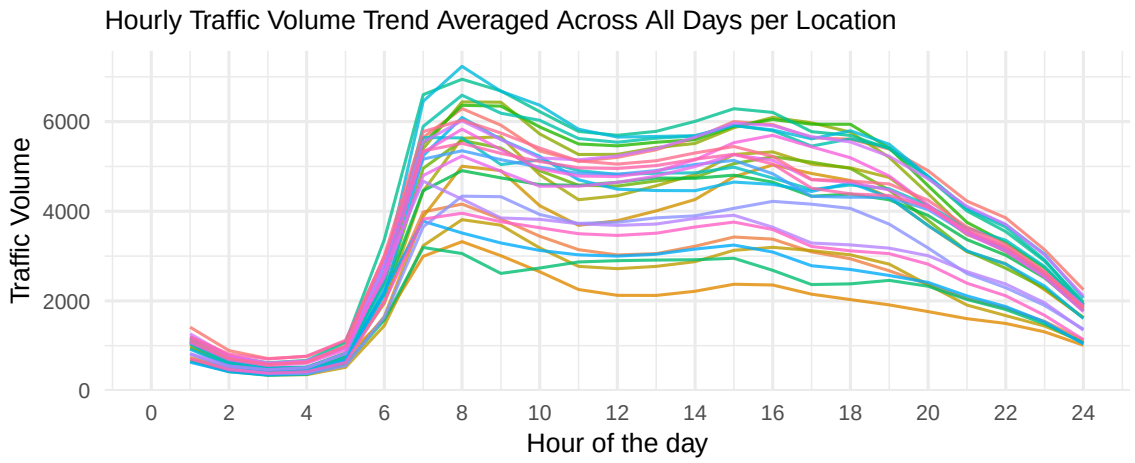

Figure 1: Hourly Traffic Volume Trend Averaged Across All Days per Location. The x-axis denotes the hour of the day, and the y-axis indicates traffic volume. Each line represents a location's average hourly traffic volume over the entire year.

overall daily traffic volume pattern, their specific traffic profiles may still differ due to location factors, such as land use and demographic factors. This suggests that when we apply dimensionality reduction on the data, it may be more appropriate to conduct it separately for each location to retain the unique representation of local temporal dynamics.

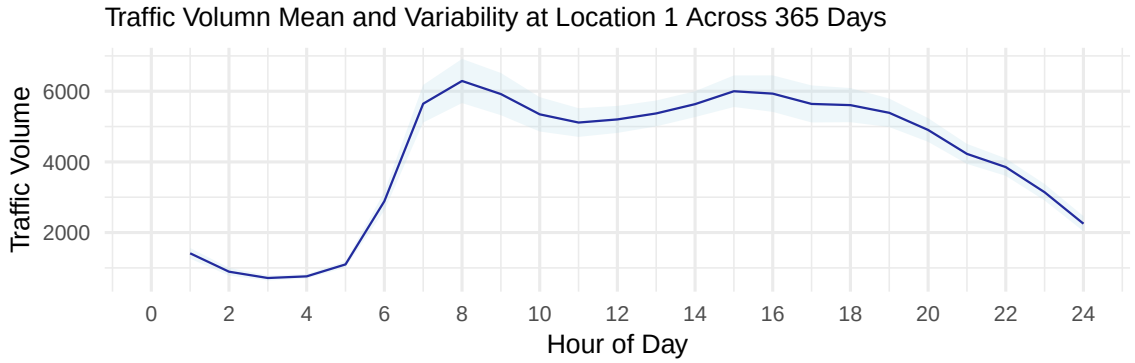

Figure 2: Traffic Volumn Pattern and Variability at Location 1. The x-axis denotes the hour of the day, and the y-axis indicates traffic volume. The solid line represents the average hourly traffic volume at Location 1 across all days of the year. The shaded region corresponds to ± 1 standard deviation from the mean, capturing the typical range of variability in hourly traffic throughout the year.

Figure 2 reveals the average daily traffic volume pattern at a single selected location, characterized by pronounced morning and afternoon peaks, which is consistent to Table 1 and Figure 1 findings. However, substantial variability in traffic volume is observed across days,

especially during daytime hours. This underscores the influence of day-to-day factors, such as weekdays versus weekends, holidays, weather conditions, and local events, on traffic volume. This imply that there are potential distinct daily traffic profiles across days.

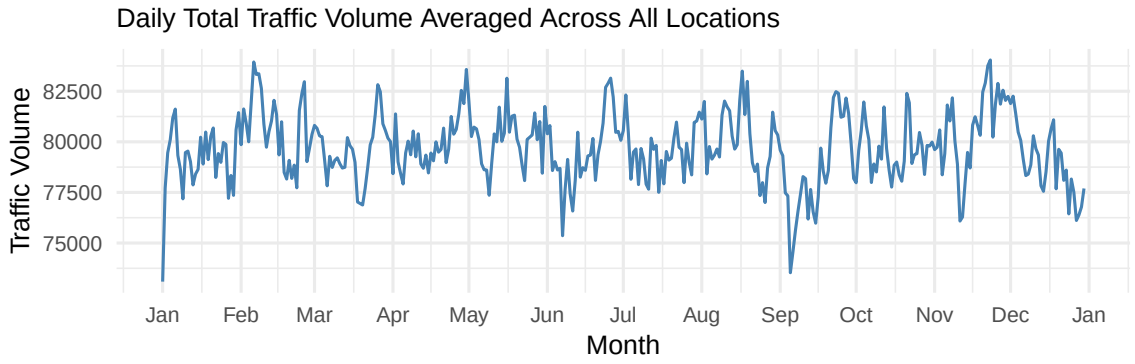

Figure 3: Daily Total Traffic Volume Averaged Across All Locations. The x-axis denotes time in months, and the y-axis represents averaged daily traffic volume across locations. Each data point on the line represents the total traffic volumes for a single day (summed over hourly data), averaged across all locations.

According to Figure 3, daily traffic volumes generally range between 77,500 and 82,500 vehicles per day throughout the year. Sharp declines occur on January 1, early June, early September, and mid-November, corresponding to major calendar events such as New Year's Day, the start of the summer school holiday, the return to school in September, and Remembrance Day on November 11 (although it is not a statutory holiday in Toronto). December also shows a gradual decline leading into the Christmas period, while the interval from June to mid-August exhibits a modest upward trend consistent with increased summer travel activity. These findings suggest that days may be meaningfully grouped into clusters, such as holiday-related low-traffic days, summer holiday days with moderately rising volumes, and stable commuting periods.

Figure 4 illustrates location-specific variations in daily traffic volume that both mirror and diverge from the overall trend. The sharp dip on January 1 is consistently observed across all locations, reinforcing its classification as a low-traffic holiday. From June to mid-August, all three locations exhibit a modest upward trend, consistent with increased summer travel activity noted in Figure 3, but the magnitude of increase vary among locations. On the other hand, December reveals more pronounced divergence. While Location 1 shows a marked de-

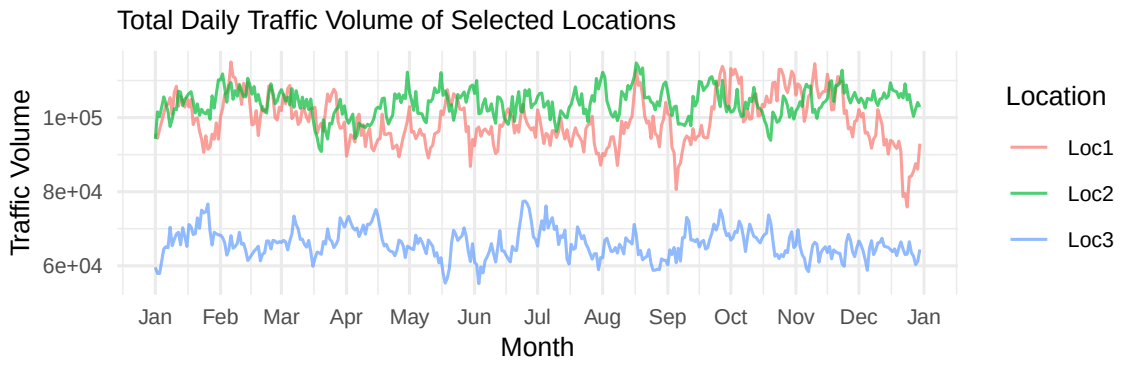

Figure 4: Total Daily Traffic Volume of Selected Locations. The figure displays daily total traffic volumes over the course of a year for three locations (Loc1, Loc3, and Loc5). The x-axis denotes time in months, and the y-axis represents total daily traffic volume. Each colored line corresponds to traffic volume trend in one location.

cline in traffic volume leading into the Christmas period, Locations 2 and 3 remain relatively stable, with Location 3 displaying a slight bump in late December. This discrepancy suggests that longer holiday periods may elicit location-dependent responses in traffic demand. While short holidays may have uniformly low traffic volume pattern cluster across locations, longer breaks and school holidays may have different magnitude of changes in traffic depending on location-specific dynamics.

3 Methodology

3.1 Principal Component Analysis (PCA)

The original dataset comprises daily traffic volume records from $\ell = 26$ locations, each measured at hourly intervals, resulting in 24 feature vectors per location. Principal Component Analysis (PCA) was conducted to project the high-resolution daily traffic volume data onto a more compact feature space, aiming to capture recurring daily structures such as morning and evening rush hour peaks. This enables efficient identification of potential clustering of traffic profiles across days. In particular, PCA was applied independently to each location-specific dataset accommodate local variability. Prior to PCA, each location matrix was standardized (i.e., centered and scaled). Although traffic volume is consistently recorded in

units of car counts, standardization was necessary due to substantial variability in hourly traffic patterns across different days within the same location (as shown in Figure 2). For each location ℓ, the number of principal components r_ℓ was chosen such that the cumulative explained variance is at least 90%. As illustrated in Figure 1, while locations exhibit broadly similar daily traffic patterns, the magnitude, timing, and shape of peaks differ. Allowing the number of principal components to vary by location enables complex sites with richer temporal dynamics to be represented with more components, while simpler profiles are captured with fewer, achieving a balance between effective dimensionality reduction and explanatory power. The 90% threshold was selected to retain sufficient detail for the downstream clustering tasks. For each location, the PCA yielded a score matrix representing the projection of the original data onto the first r_ℓ principle components. These score matrices were then extracted and concatenated into a single matrix, forming a unified representation of the most informative features across all locations.

3.2 Uniform Manifold Approximation and Projection (UMAP)

To explore latent structure in daily traffic profiles, we applied Uniform Manifold Approximation and Projection (UMAP) to the combined feature matrix of 365 daily observations across sum of r_ℓ dimensions. UMAP is a nonlinear dimensionality reduction technique that aims to preserve local and global structure by constructing a high-dimensional graph and optimizing its low-dimensional projection so that the graphs in the two dimensional space are "similar", and it is effective for identifying clusters in the data. We varied two key parameters, number of nearest neighbors and target embedding dimensions, to explore how different configurations affect the clustering results. Specifically, we tested on combinations of 5, 10, and 15 neighbors with 2 UMAP components. The number of neighbors controls the balance between local sensitivity and global coherence, while the number of UMAP components determines the dimensionality of the output projection space. We chose 2 as the target embedding dimension to facilitate visualizations for identification of potential clusters on a 2D plane. Results from each UMAP configuration were visualized to assess whether distinct clusters emerged. We selected the configuration that produced more than one clus-

ter, as a single undifferentiated grouping would suggest that daily traffic patterns are not temporally distinguishable. For the chosen projection, we examined the data points (i.e., individual dates) within each cluster to evaluate their temporal coherence, focusing on their calendar month and intra-month phase (early, mid, late). This analysis allows us to assess whether traffic patterns across locations exhibit seasonal regularities, such as holidays, summer periods, etc.

3.3 Statistical Software

All statistical analyses were conducted in R version 4.3.0 (R Core Team 2023). We relied on several R packages to support data cleaning, modeling, and visualization, including: openxlsx (Schauberger and Walker 2025), tidyr (Wickham, Vaughan, and Girlich 2024), dplyr (Wickham et al. 2023), knitr (Xie 2025) for data manipulation, umap (Konopka 2023) for analysis, and ggplot2 (Wickham 2016), viridis (Garnier et al. 2024), ggfortify (Horikoshi and Tang 2018), plotly (Sievert 2020), gridExtra (Auguie 2017), cowplot (Wilke 2025) bfor data visualization and generating tables.

4 Results

PCA was applied independently to each of the 26 location-specific datasets, and the number of retained components, r_ℓ, ranges from 2 to 9, with a median of 7, reflecting differences in temporal complexity across locations. The resulting retained components captured 90.08% - 93.55% of the cumulative variance across locations. Total retained components is 164.

Principal components can be interpreted in terms of meaningful traffic dynamics. As an example, we showcase the principal components obtained from Location 7. This location retains two components to explain 90% of the total variance in daily traffic volume profiles. After projecting the original data into the PCA feature space, we selected extreme observations along each principal component axis to investigate their interpretability. The chosen days are: Day 5, 222, 267 and 363 (see Figure 8). Day 5 and Day 222 exhibit extreme

positive and negative score values on PC1, respectively, while maintaining near-zero score values on PC2. Day 267 and Day 363 show extreme positive and negative score values on PC2, with minimal difference along PC1.

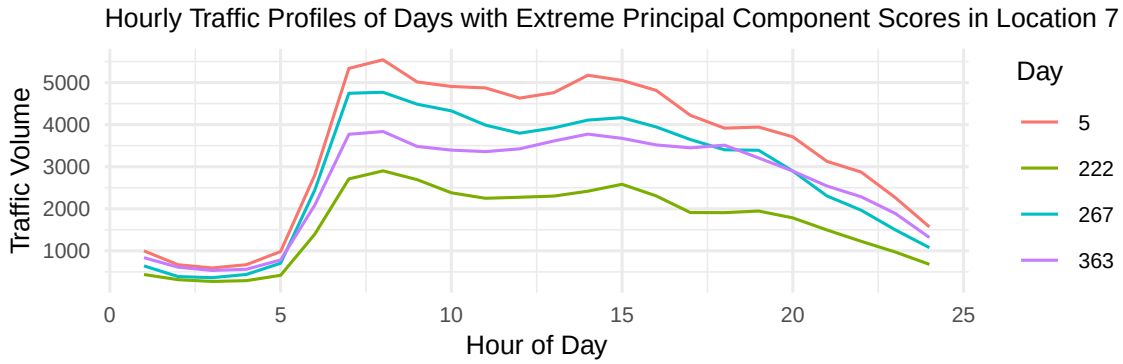

Figure 5: Hourly Traffic Profiles of Days with Extreme Principal Component Scores in Location 7. The x-axis denotes hour of day, and the y-axis represents traffic volume. Each line corresponds to the traffic volume trend on a specific day at Location 7.

Figure 5 illustrates how the first two principal components capture distinct type of variation in location 7's daily traffic pattern. PC1 appears to encode the overall magnitude of traffic volume since Day 5, with a highly positive PC1 score, shows elevated traffic throughout the day, while Day 222, with a strongly negative PC1 score, exhibits consistently low volume. In contrast, PC2 may reflect the position of midday dip, where negative values correspond to earlier peaks and positive values to later ones. Day 267, with a positive PC2 score, shows a delayed midday dip, whereas Day 363, with a negative PC2 score, displays an earlier dip.

The UMAP matrix, concatenated by combining all location-specific PCA score matrices, has dimension 365×164. Among the tested UMAP configurations (i.e., 5, 10, and 15 neighbors with 2 UMAP components), only the setting with 5 neighbours yielded more than one distinct cluster. This suggests that finer local sensitivity is necessary to reveal meaningful variation in daily traffic profiles.

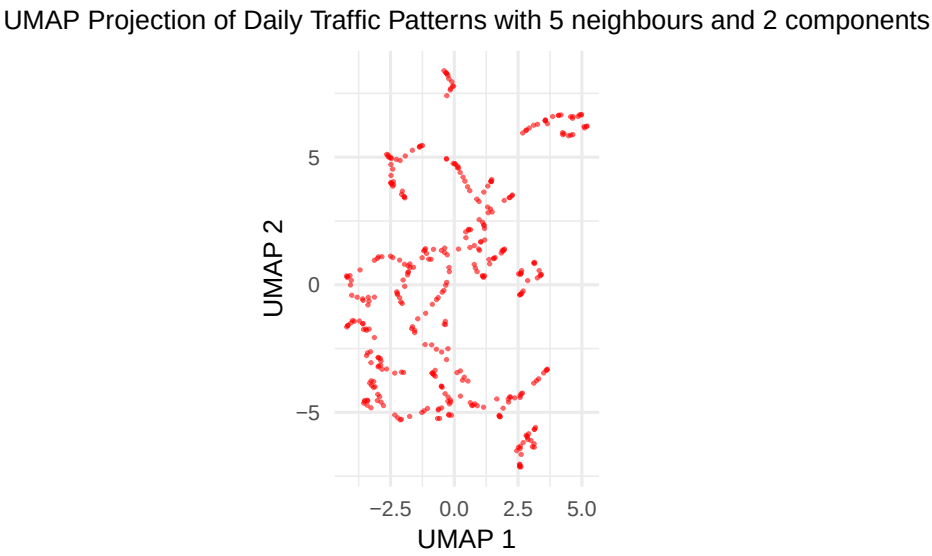

Figure 6: UMAP Projection of Daily Traffic Patterns with 5 neighbours and 2 components. Each point in the plot represents a single day's traffic pattern across locations, projected into a two-dimensional space using UMAP.

According to Figure 6, the projected data points formed some visually separable groupings, indicating that certain dates share similar traffic patterns across locations. For instance, there are clusters on top right corner, top center, upper left, bottom right corner.

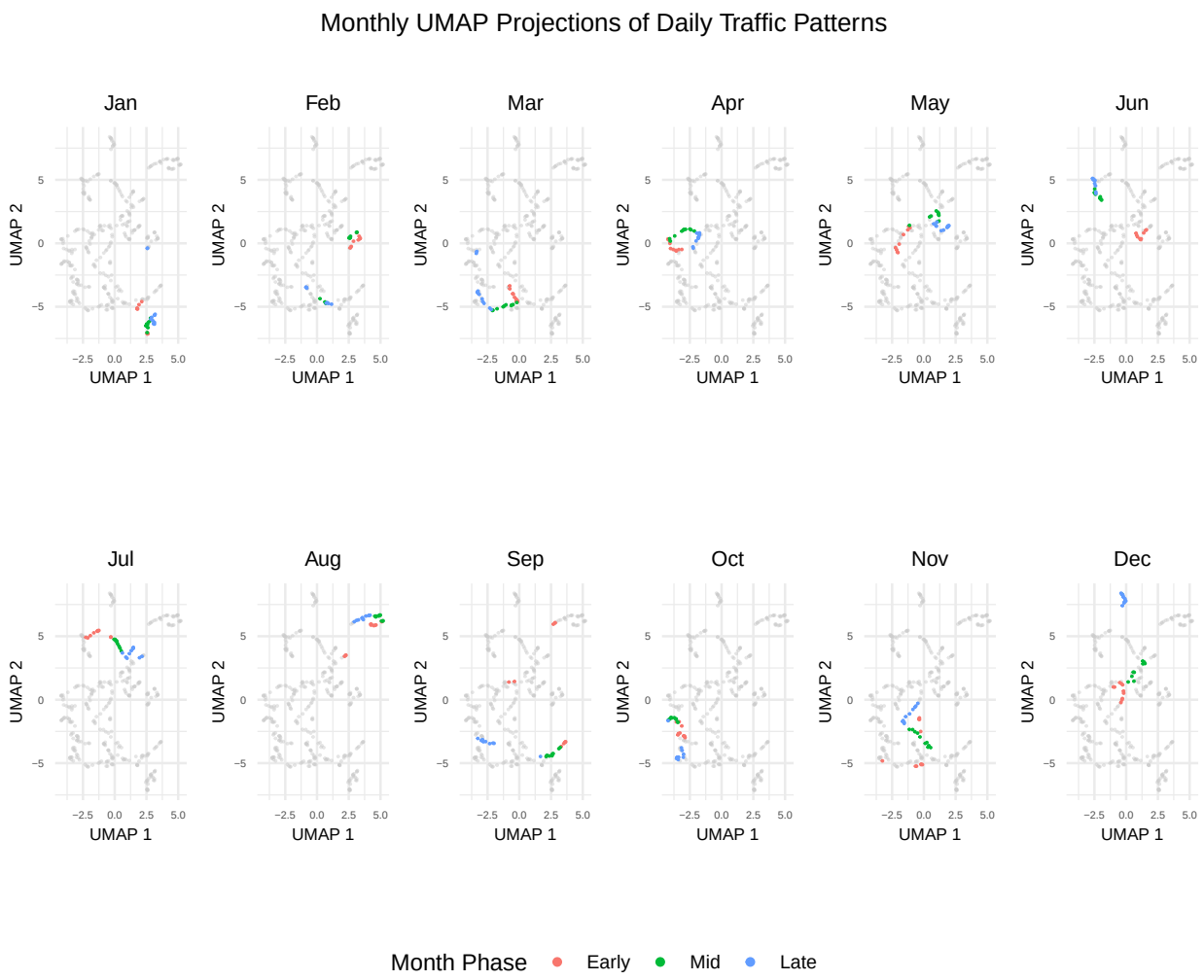

Figure 7: Monthly UMAP Projections of Daily Traffic Patterns. Each panel displays a two-dimensional UMAP projection of daily traffic profiles, with all 365 days shown in gray for context. Within each panel, only the dates belonging to the corresponding calendar month are highlighted in color according to their intra-month phase: early, mid, or late. Early corresponds to day 1-10 of the month, mid corresponds to day 11-20 of the month, late corresponds to day 21 to the last day of the month.

According to Figure 7, several clusters in the UMAP projection correspond to distinct calendar periods, suggesting recurring traffic regimes. We observe that mid-June, late June, and early July dates formed the cluster in the upper-left quadrant. This cluster may capture the transition into summer vacation, when school closures and shifting work schedules lead to reduced peak-hour congestion and increased midday travel. Similarly, August days also form a cluster at the top-right corner, which reflects late-summer travel dynamics, such as lower peak-hour congestion. Since August days are not clustered with the early summer holiday

days, this suggests that traffic behavior vary between early and late summer holiday. Moreover, the cluster at the center-top region contains points from late December, potentially reflecting similar traffic behavior around the Christmas holiday. During Christmas season, schools are closed and many people take time off work, people also tend to stay home to celebrate Chirstmas and travel less, leading to a decrease in traffic volumes across locations and throughout the day. This uniform reduction across time of day produces a distinct traffic profiles, which UMAP captures as cluster of days with low activity throughout the day. On the other hand, the bottom-right cluster is dominated by mid and late January days, likely reflects the gradual return to routine traffic behavior following the Christmas and New Year holidays. During this transitional period, the morning and evening peak hours begin to re-emerge, but the overall traffic volume may still remain lower than typical non-holiday day, potentially due to extended vacations or delayed re-openings in certain workplaces. As a result, daily traffic profiles during this phase forms a distinct cluster in the UMAP space, which is separated from both holiday and fully normalized non-holiday traffic patterns. In contrast to the distinct clusters observed around holidays and seasonal transitions, the remaining days form a broad, continuous block that extends diagonally from the bottom-left to the upper-right of the UMAP embedding. This region corresponds to days from late February to mid June and September to November, which are months typically characterized by regular weekday-weekend traffic patterns with stable commuting routines, consistent school schedules, and minimal weather disruptions. The spatial continuity of this block suggests that these days share similar daily traffic structures, with the usual peak hours and predictable flow distributions across locations. These findings demonstrate that the UMAP projection clusters capture some meaningful structure in daily traffic behavior, distinguishing between holiday-driven deviations, post-holiday transitional periods, and routine commuting regimes.

5 Discussion

5.1 Key findings

In this study, PCA is applied to daily traffic volume profiles measured at hourly intervals, reducing each 24-dimensional time series to a compact set of 2 to 9 principal components per location while retaining at least 90% of the variance in the original data. This suggests that location-specific hourly traffic volume may be shaped by some consistent temporal structure rather than random fluctuations each day. Using these location-specific PCA-reduced representations of daily traffic volume time series, UMAP results in some clustering of dates that are deemed to have similar traffic volume behavior across the calendar year. The distinct UMAP clusters corresponding to summer holidays, the Christmas season, and post-holiday transitions suggest that traffic volumes and peak-hour dynamics shift in consistent, calendar-driven ways. These insights have practical implications for transportation planning, demand forecasting, and infrastructure maintenance schedule. For instance, public transit authorities might adjust service frequency to match reduced demand during school closures in early-summer and increase capacity during late-summer travel periods. Maintenance crews can schedule disruptive roadwork during low-activity clusters closer to Christmas season or mid-August, minimizing impact on commuters. With these insights, agencies can better anticipate fluctuations in roadway usage and tailor interventions accordingly.

5.2 Limitations

First, the interpretation of UMAP components remains inherently abstract, as the axes do not correspond to specific physical or temporal features, making direct attribution challenging. The distance between clusters are also not interpretable, i.e., pair of clusters that are farer away do not necessarily mean their underlying traffic patterns differ more. There is no numerical method to quantify extent of the contextual difference between clusters. Second, cluster identification in this study was mainly based on visual inspection, which may be subjective. While several groupings appeared visually distinct, no formal statistical test was

applied to validate their separability. Moreover, the resulting cluster structure is sensitive to the UMAP configuration. Increasing the number of neighbors (e.g.,10 or 15, see Figure 9) led to a more homogenized projection, with all data points forming a single cluster. One could argue that there is no distinct traffic pattern clusters. Consequently, the conclusions drawn about which days exhibit similar traffic patterns, which depends on the resulting cluster structure, is sensitive to the choice of UMAP configuration parameters, in particular the chosen number of neighbors. These limitations suggest that while UMAP offers a useful exploratory tool to identify clusters in daily traffic patterns, further validation using complementary methods should be considered. Lastly, this study relies on a synthetic dataset, and we lack access to the true calendar dates and geographic identifiers. Each day is labeled generically (e.g., day1, day2) and each location is anonymized (e.g., loc1, loc2) in the data. So, we cannot confirm whether day1 corresponds to January 1, nor can we determine the year or region represented by each location. This limits our ability to validate calendar-based interpretations or connect observed traffic regimes to external contextual data such as holidays, weather, or school schedules. While the temporal ordering of days appears consistent, the absence of metadata restricts the precision of behavioral inferences and the generalizability of findings.

5.3 Next steps

We should consider incorporating real-world metadata, including true calendar dates, geographic identifiers, and contextual variables (e.g., weather and school schedules), which would enable more precise validation of the current study's interpretations of clusters. Besides, we can replicate the analysis on an expanded dataset which includes traffic data covering multiple years, this allows for the validation of the cluster structures identified in current study and incorporation of inter-year variability in traffic volume, enhancing the generalizability of our findings. In this study, we aggregated traffic volume data from 5-minute intervals into hourly totals prior to applying PCA and UMAP. Future research may benefit from retaining the original temporal resolution to further capture the fine-grained fluctuations in traffic patterns.

6 References

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Boyce, David. 2012. "Predicting road traffic route flows uniquely for urban transportation planning." *Stud. Reg. Sci.* 42 (1): 77–91.

Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, et al. 2024. *viridis(Lite) - Colorblind-Friendly Color Maps for r*. https://doi.org/10.5281/zenodo.4679423.

Horikoshi, Masaaki, and Yuan Tang. 2018. *ggfortify: Data Visualization Tools for Statistical Analysis Results*. https://CRAN.R-project.org/package=ggfortify.

Konopka, Tomasz. 2023. *umap: Uniform Manifold Approximation and Projection*. https://CRAN.R-project.org/package=umap.

Ma, Tao, Fang Yao, and Zhou Zhou. 2024. "Network-level traffic flow prediction: Functional time series vs. functional neural network approach." *Ann. Appl. Stat.* 18 (1).

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Schauberger, Philipp, and Alexander Walker. 2025. *openxlsx: Read, Write and Edit xlsx Files*. https://CRAN.R-project.org/package=openxlsx.

Sievert, Carson. 2020. *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman; Hall/CRC. https://plotly-r.com.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2024. *tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Wilke, Claus O. 2025. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*. https://CRAN.R-project.org/package=cowplot.

Xie, Yihui. 2025. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.

A Appendix

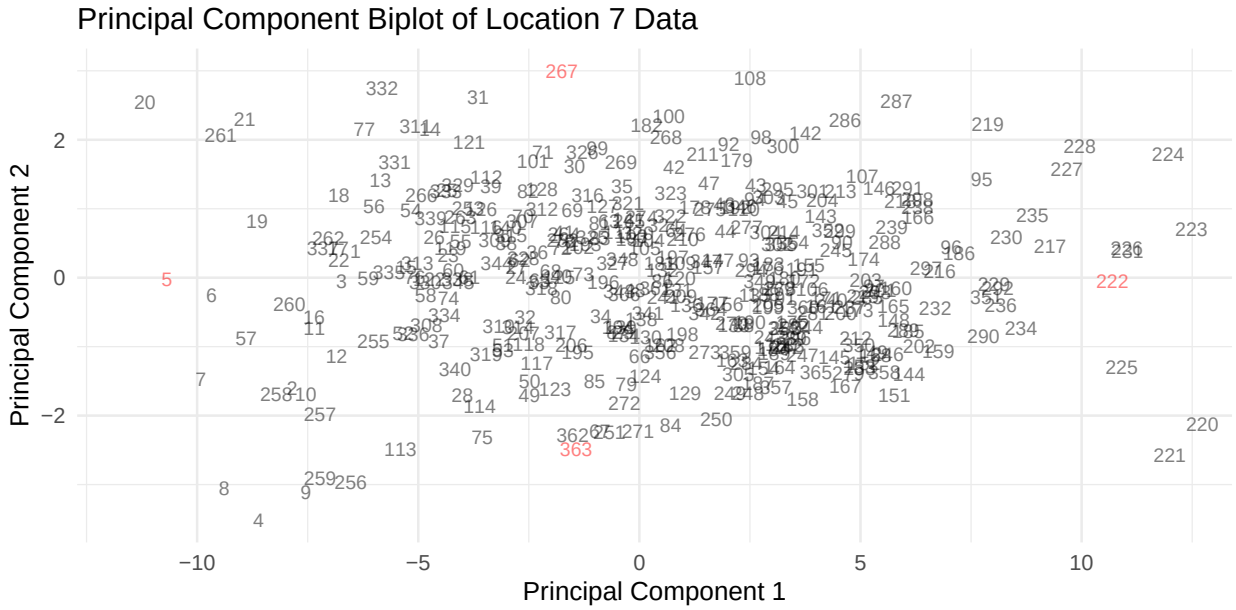

Figure 8: Principal Component Biplot of Location 7 Data. Each point represents the projection of a traffic volume pattern at a day at Location 7. The number corresponds to the day number. Extreme days are denoted in red.

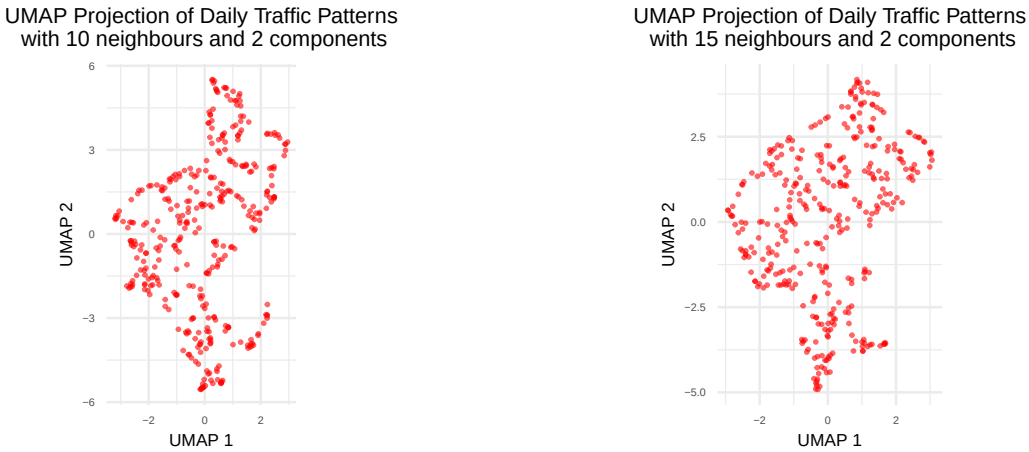

Figure 9: UMAP projections of daily traffic patterns using 10 (left) and 15 (right) neighbours and 2 components. Each point in the plot represents a single day's traffic pattern across locations, projected into a two-dimensional space using UMAP.