Multivariate Routes Through Traffic Anomalies

Erin Xu

December 1st, 2025

Introduction

The inevitability of traffic congestion in urban locations implies that studying its patterns and underlying dynamics can enable daily commuters and transportation authorities to transition from reactive management to proactive intervention, which may lead to overall reduced congestion, lower emissions and improved commuter safety around high-density areas. However, accurately detecting and predicting traffic anomalies that cause significant delays remains a challenge due to the inherent complexity of traffic dynamics, which are continuous, stochastic, spatiotemporally autocorrelated and cross-correlated (Columbia University Mailman School of Public Health, n.d.).

As existing prediction has evolved from interval-based pointwise in univariate time-series data to a functional approach at the network-level utilizing neural networks (Ma et al., 2024), this project offers a comparative assessment of common multivariate statistical techniques that are highly interpretable as benchmarks for further study. The goal of this paper is to identify and classify location-specific anomalies from the intraday patterns of traffic volume flow collected across 26 monitoring sites around the University of Toronto by comparing principal component analysis (PCA), factor analysis (FA), and independent component analysis (ICA) methods, which are selected for their ability to achieve dimension reduction of spatio-temporal data, interpret latent regimes to reveal original causes of anomalous traffic volumes at each location, and isolate mixed data components.

Data Description

The dataset, synthetically modeled from Ma et al. (2024), has a natural tensor structure consisting of 26 locations (l), 384 (n) days each, and 288 (p) five-minute time points per day. Then each slice $X \in \mathbb{R}^{n \times p}$, called the daily traffic matrix, corresponds to one location and forms a 384×288 matrix, with each entry as volume in vehicles.

No data cleaning was required, as all locations share uniform dimensions and contain no missing observations. Exploratory visualizations were produced for two randomly selected locations to conserve space. Spaghetti plots by location and by day, accompanied by summary statistics, highlight clear daily peak structures. The first location exhibits lower median

flow than the second, suggesting an area with less traffic, perhaps a more residential area. In contrast, the two randomly sampled days display similar median volumes, consistent with typical weekday patterns. These preliminary observations motivate the subsequent use of statistical methods to quantify temporal and spatial structure in the full dataset. A method that would fully reveal the context of traffic volume with location should involve identification of each location's address.

The coefficient of variation (CV) and autocorrelation function (ACF) analyses indicate that the traffic system exhibits a strong diurnal rhythm and persistent temporal dependence, supporting the use of dimension-reduction methods. CV quantifies day-to-day variability at each 5-minute interval. Averaged across locations, the CV curve shows highest variability overnight, a sharp decline during the morning, and a pronounced minimum around midday, followed by increasing variability toward the evening peak. The narrow CV shadow around noon suggests highly consistent midday traffic, whereas the wider shadow during peak hours reflects differing commuter patterns across locations.

Temporal autocorrelation was assessed at a representative site (Location 5), selected because its mean CV is closest to the median across all locations. The ACF reveals strong short-term persistence: autocorrelation decays gradually over several hours and remains positive even at four-hour lags, implying that intraday traffic evolves smoothly. The absence of negative correlations indicates a stable daily cycle.

The spatial correlation heatmap has shows that the center locations generally are correlated with each other, which suggests functional subregions where locations follow similar demand cycles. Most correlations fall in the moderate positive range, indicating coordinated but heterogeneous behavior across the network. Therefore location-specific anomaly detection is meaningful with location specific nuances, and network-wide patterns could also be coherent enough as the system has shared temporal dynamics.
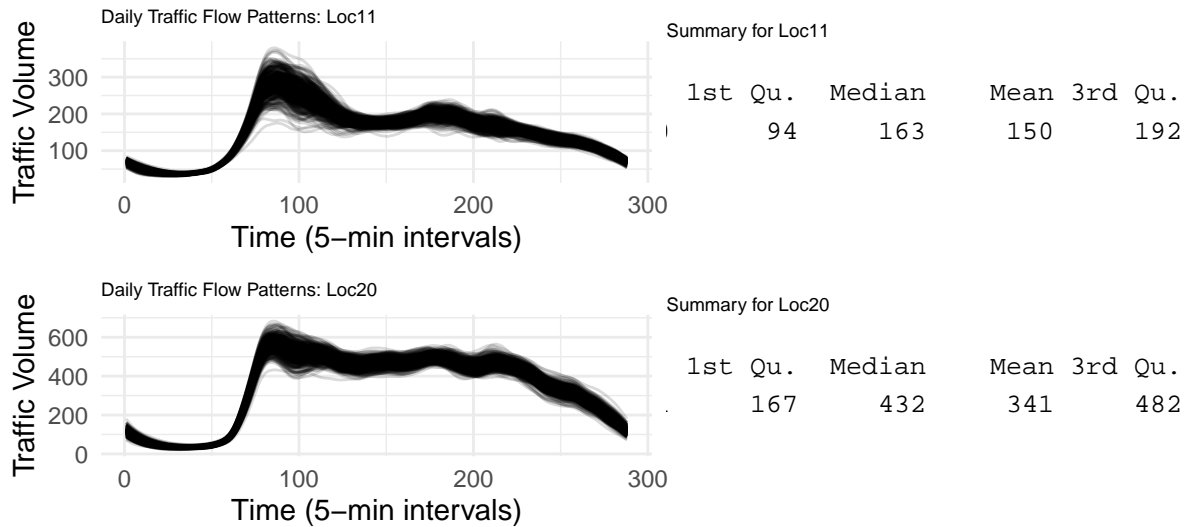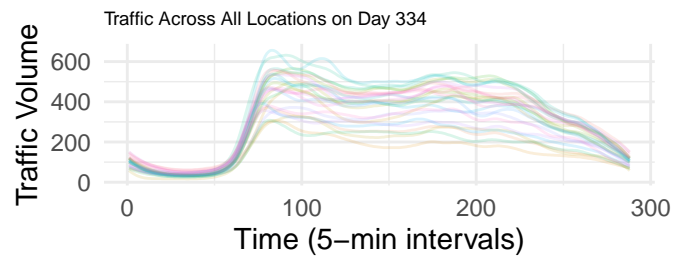

Figure 1: Spaghetti Plot by Location

Summary for Day 334

1st Qu.	Median	Mean	3rd Qu.
137.0	299.0	284.9	423.0

Traffic Across All Locations on Day 334

Summary for Day 199

1st Qu.	Median	Mean	3rd Qu.
132.0	274.0	270.6	403.0

Traffic Across All Locations on Day 199
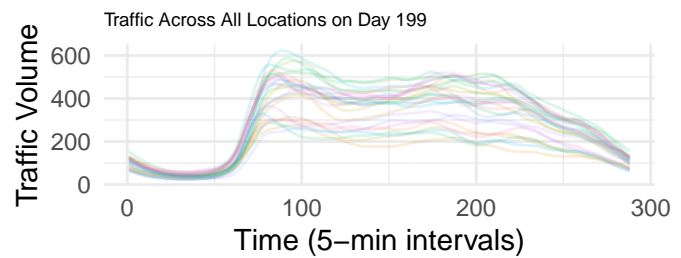
Figure 2: Spaghetti Plot by Day

Traffic Variability Throughout the Day

Coefficient of Variation (CV = SD / Mean) across all locations

Temporal Autocorrelation (ACF): Loc5
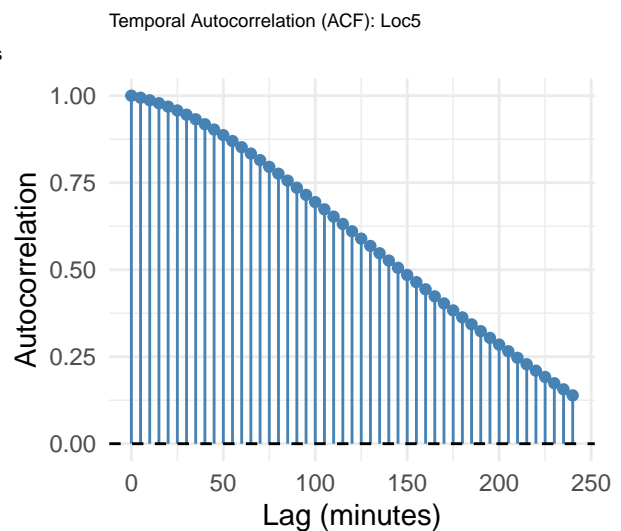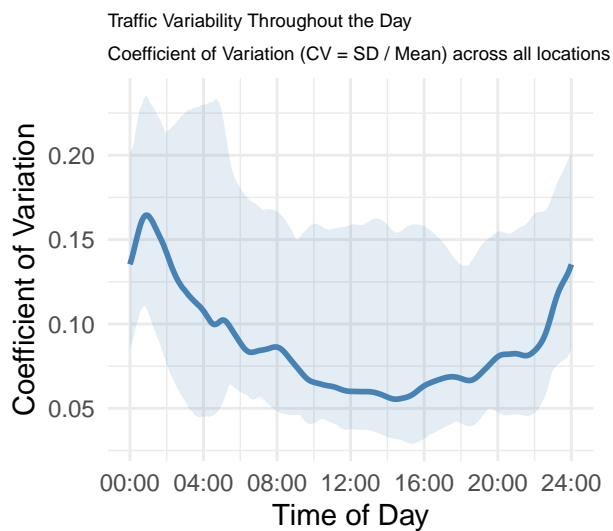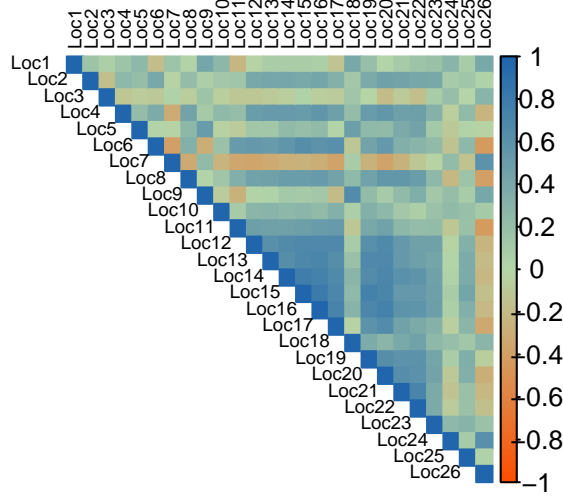
Figure 3: CV and ACF

Figure 4: Spatial Correlation: Daily Total Traffic Across 26 Locations

Methodology

The traffic dataset is high-dimensional, functional in nature (with $p = 288$ time points per day), and exhibits substantial temporal dependence.

Consequently, dimension reduction constitutes a fundamental component of the anomaly detection framework with PCA serving as the primary dimension-reduction mechanism, as each daily traffic profile is represented by a matrix $X \in \mathbb{R}^{n \times p}$, and PCA assumes that although each curve x_i lies in a high-dimensional ambient space $\mathbb{R}^p$, its intrinsic variation is concentrated on a low-dimensional manifold. This representation is consistent with standard functional data analysis practice, where curves are expressed in a reduced basis and where PCA provides a computationally efficient empirical basis for large-scale datasets.

For each location, the 384×288 traffic matrix X was transposed so that rows corresponded to days and columns to time points. PCA was then conducted on the centered traffic matrix, $X_c = X - \mathbf{1}_n \bar{x}^\mathsf{T}$, where $\bar{x}$ denotes the sample mean profile. Let $V = [v_1, ..., v_p]$ denote the orthonormal loading functions (the empirical eigenbasis), and let $S = X_c V$ denote the corresponding score matrix. The score of day i on component j is given by $s_{ij} = X_{c,i}^\mathsf{T} v_j$. The number of retained components k was selected via the 90% variance-explained criterion: $\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \geq 0.90$, where $\lambda_j = \sigma_j^2$ denotes the variance explained by component j, which can be obtained through the squared singular values of the sample covariance matrix. This process highlights major regimes, like a morning peak, and discards higher-frequency noise.

Then anomalies were identified directly from the PCA score matrix $S \in \mathbb{R}^{n \times k}$ as PCA scores are generally computed via singular value decomposition (SVD) of each x_i and are uncorrelated and form an orthogonal basis. For each component j, a day i was flagged as anomalous if $s_{ij} < Q_{1,j} - 1.5\,\mathrm{IQR}_j$ or $s_{ij} > Q_{3,j} + 1.5\,\mathrm{IQR}_j$, where $Q_{1,j}$, $Q_{3,j}$, and IQR_j denotes the first quartile, third quartile, and interquartile range of component j respectively.

4

In comparison to Mahalanobis-distance methods, this boxplot method is robust enough to not assume normality and is standard in functional data outlier detection, which is important as PC scores are not guaranteed to be multivariate normally distributed (MVN), although surely uncorrelated. Boxplots give interpretable, location-specific anomaly sets that can be classified by time-of-day (Shang & Hyndman, 2010).

FA provides an alternative dimension-reduction framework. Intuitively, FA attempts to explain the covariance structure of x_i by assuming that the observed factors are linear combinations of latent factors with an unique error term. In other words, when there is correlation in x_i, it is due to the influence of common latent variables that can be uncovered. Mathematically, this means that an observed random vector x satisfies $x = \Lambda z + \varepsilon$, where z is a lower-dimensional latent vector, Λ is a loading matrix (factor scores and loadings), and ε represents idiosyncratic noise. FA requires a full-rank, invertible sample covariance matrix; estimation therefore proceeds via iterated PCA and not on the raw data, until uniqueness variances converge. Both FA and PCA reduce the dimension of the data. In contrast, FA may provide more interpretable loadings in comparison to PCA, as latent factors are viewed as broad concepts or ideas that contribute to a pheonmenon, and these are highly valuable in the social sciences - there is meaning all around.

FA anomalies were also detected using the same $1.5 \times IQR$ rule applied to the factor score matrix F, where $f_{ij} < Q_{1,j} - 1.5\,\mathrm{IQR}_j$ or $f_{ij} > Q_{3,j} + 1.5\,\mathrm{IQR}_j$. Because FA captures deviations from latent structural factors rather than maximizing total variance, FA anomalies could correspond to days whose patterns violate the inferred latent structure like shifted peaks, whereas PCA anomalies reflect variance-aligned distortions like spikes associated with incidents.

ICA was applied to the PCA scores to extract statistically independent latent signals embedded within the traffic profiles. Intuitively, ICA identifies original components (inherent cause) as linear combinations for each x_i, beyond a mathematically rigorous uncovering of components (loadings) like in FA or PCA. While those loadings may certainly exist, they may not be source signals. Therefore, ICA assumes whitened inputs with identity covariance and therefore requires PCA preprocessing, like in FA (Hyvärinen & Oja, 2000). Then, it mathematically follows that ICA yields the decomposition $S_{\mathrm{PCA}} = A S_{\mathrm{ICA}}$, where S_{ICA} contains the source signals (scores) and A is the mixing matrix (loadings). Unlike PCA and FA, ICA uses information on statistical independence and non-gaussianity to recover source signals, meaning that it doesn't rely on the limited information in the covariance data matrix. It is reasonable to assume data is non-gaussian, as while overall traffic sensors take in a mixture of data coming from various components and by the central limit theorem, any mixture of components will be more gaussian than the components themselves, but PC scores (in this case, original components) themselves are not guaranteed to be multivariate normally distributed (MVN), that is to say, non-gaussian. Then, ICA's goal is to maximize the non-gaussianity of PC scores (components) - this is the projection pursuit. ICA does not struggle with the factor rotation problem that appears in FA.

ICA anomalies were detected via the same $1.5 \times \mathrm{IQR}$ rule applied to the independent source scores. Intuitively, ICA anomalies represent rare independent micro-events that sharply distort the traffic curve of a different location, for example a sudden dip/spike that only

lasts a few intervals that PCA smooths out, or odd jumps that FA distributes across factors and randomness. ICA is able to provide true causal analysis.

All anomalies identified across PCA, FA, and ICA were projected onto their two-dimensional principal component subspace and partitioned using k-means clustering (Piech & Ng, 2013). This method was selected for its computational efficiency and suitability for continuous Euclidean feature spaces. The number of clusters was chosen by maximizing the average silhouette coefficient: $k^* = \arg\max_k \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \right\}$, where $a(i)$ is the average within-cluster distance and $b(i)$ is the minimum average between-cluster distance (Rousseeuw, 1987). This procedure yields interpretable groups of anomalies that reflect distinct structural perturbations in daily traffic dynamics. Taken together, PCA, FA, and ICA form a benchmark set: PCA captures global variance-driven deviations, FA captures structural inconsistencies relative to latent factors, and ICA captures independent localized perturbations that correspond to the original sources of anomalous traffic volume. Using all three offers a comprehensive view of anomalous behavior from orthogonal interpretive perspectives: variance, latent structure, and independence. This ensures that anomalies detected are robust to modeling assumptions and interpretable in terms of their functional, temporal, and structural characteristics.

Results

PCA retained between 2 and 9 components across all 26 locations, consistently explaining at least 90% of total variance, with cumulative explained variance ranging from 0.9006 to 0.9404.

PCA detected between 4 and 31 anomalous days per location, with most locations falling in the 15–25 anomaly range. This demonstrates that although each daily curve contains 288 five-minute observations, the dominant variation is always low-dimensional.

Across locations, the first two principal components (PCs) followed consistent functional interpretations:

1. PC1 (global amplitude): Loadings were strictly non-negative and broadly elevated across the 24-hour period, indicating that PC1 captures overall daily traffic volume, such as comparing high-traffic days against quiet days.

2. PC2 (shape distortion or timing shift): PC2 loadings exhibited a morning–evening contrast, with positive loadings during morning peaks and negative loadings in the evening (or vice versa). This component reflects differences in peak timing, where anomalously shifted or flattened peaks appear as extreme PC2 score values.

3. Higher order PCs (3-9): They would continue to capture localized deviations.

Intuitively, these anomalies correspond to days with unusually large or small PC scores relative to the interquartile range for at least one retained component. Because PCA aligns
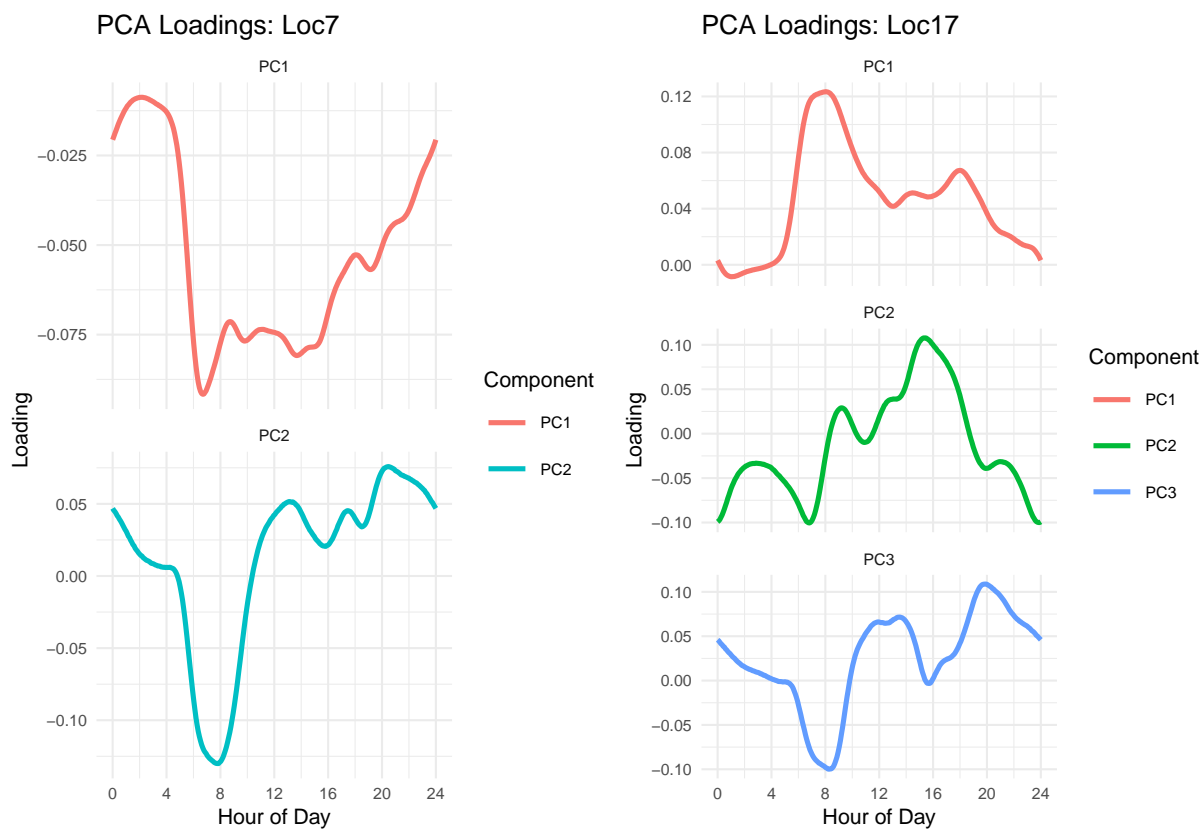
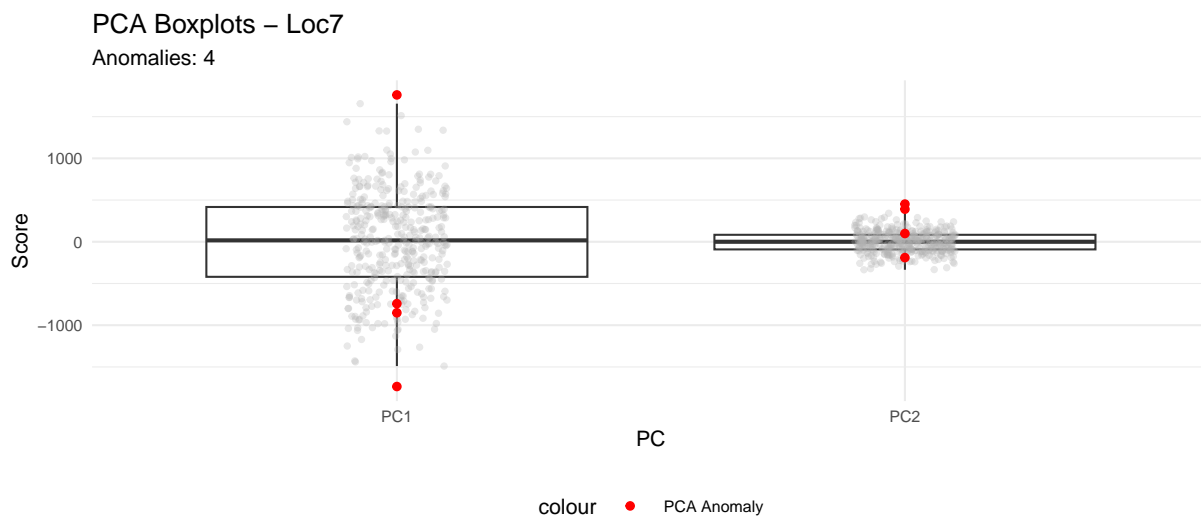Figure 5: PCA Loadings for 2 Locations with Anomalies

Figure 6: Boxplot for 1 PCA Location with Anomalies

with variance-maximizing directions, these anomalies could represent large-scale distortions such as full-day surges, holiday traffic patterns, or major disruptions.

Factor Analysis successfully converged for 17 of 26 locations, always extracting 3 factors. FA identified between 5-16 anomalous days at successful locations. From the output messages, convergence failures usually occured when PCA dimensionality was too low (2-3).
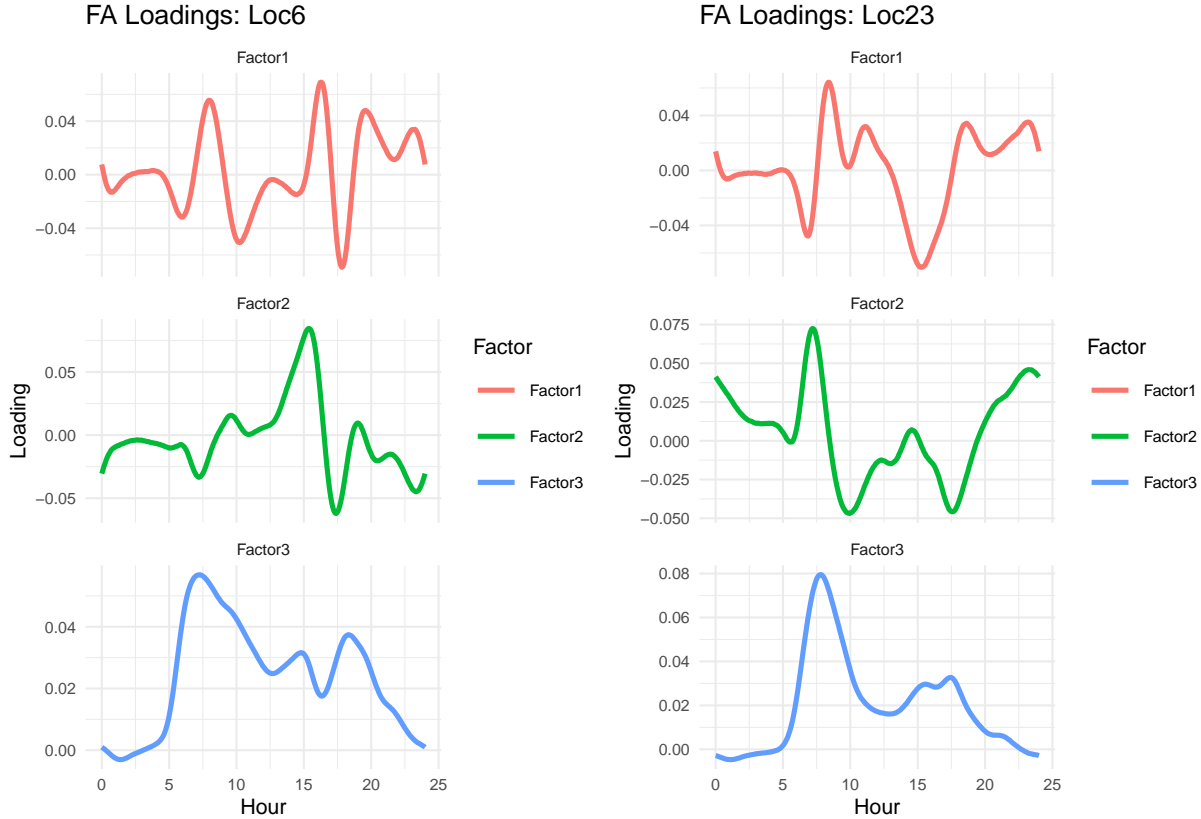

Figure 7: FA Loadings for 2 Locations with Anomalies

Across locations, the first FA loadings followed these interpretations:

1. FA1 (baseline daily volume)

2. FA2, FA3 (shape distortions): They are not variance-maximizing directions.

Intuitively, these anomalies represent days that deviate from the latent factor structure, rather than simply from variance. While PCA anomalies capture overall surges or large-scale changes, FA anomalies identify violations of latent behavioral patterns. That could be peak shifting within latent factors, inconsistent "shape modes" of daily traffic, or breakdowns of the usual relationships between morning and afternoon traffic. Because of the reduced anomaly identification, FA is more conservative and structurally selective.
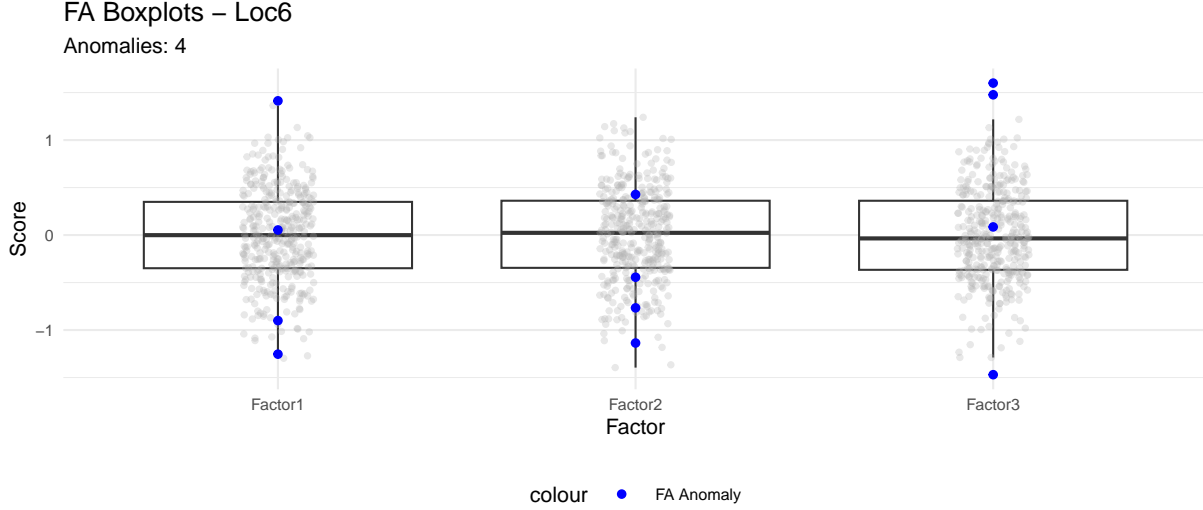
Figure 8: Boxplot for 1 FA Location with Anomalies

FA score distributions appear more compact, with fewer extreme outliers. FA captures deviations in latent covariance structure (e.g., changes in the relationship between morning and evening peaks) rather than raw amplitude. The narrow score boxes suggest stable relationships between their latent factors across days. When FA does produce outliers, they appear as isolated points, indicating rare disruptions in factor relationships. This narrower spread explains why FA produces moderate anomaly counts—it flags structural departures, not amplitude-driven ones.

ICA converged successfully at all 26 locations, producing 2–3 independent components depending on PCA dimensionality as expected after PCA whitening. ICA detected between 3 and 22 anomalies, typically between 8 and 15 anomalies per location.

Across all locations, the first sources followed these functional interpretations:

1. ICA Source 1 (localized spikes): The shape is a sharp positive or negative jump over a very short interval (10–30 minutes). This could mean a sudden incident, brief surge, or lane closure, or short-lived dip (sometimes negative spike). It is statistically independent from the smooth daily curve.

2. ICA Source 2 (independent morning distruption): The is a strong swing isolated around around 7–10 AM. This could be interpreted as morning congestion anomalies school-day vs. holiday pattern, morning-only incident, or weather-related delays. These don't necessarily correlate with evening patterns as they are independent from other components.

3. ICA Source 3 (independent evening disruption): A bump/dip concentrated around 4–7 PM can be interpreted as evening rush anomalies, evening event traffic, or some end-of-day incident. ICA finds this separate from morning events because morning/evening anomalies often do not co-occur.
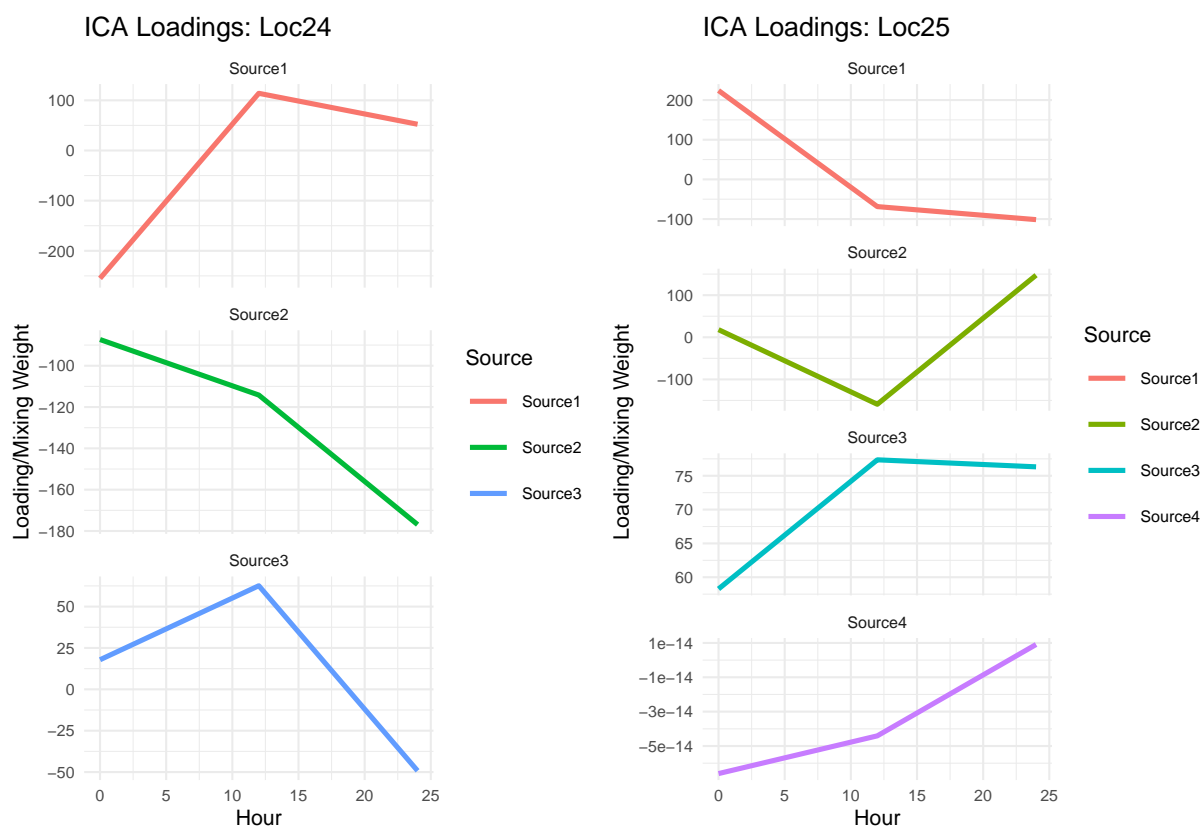
9

Figure 9: ICA Loadings for 2 Locations with Anomalies

Intuitively, ICA highlights micro-structure anomalies that PCA smooths out and FA diffuses across latent factors. ICA anomalies often overlapped with PCA and FA anomalies, but identified unique, fine-grained disturbances, consistent with ICA's sensitivity to localized, high-frequency deviations. These could be: abrupt, short-lived spikes, dips at specific time intervals, independent transient disruptions (perhaps weather or accidents). In observation, university traffic is regular but shaped by overlapping processes, ICA is uniquely suited to separate them into interpretable underlying signals. Concretely, ICA would reveal class changeover cycles (large groups exiting lecture halls on the hour, groups commuting to specific locations), residence-to-campus migration patterns in the morning, library closing times, sports events. Then, anomalies could be irregular pedestrian surges on certain days like the sports events, road closures, and students going to exams. These signals exist simultaneously and overlap; PCA cannot fully separate them because PCA components remain orthogonal but not independent. ICA isolates each hidden source and provides the most rigorous identificiation of anomalies by pre-existing cause.
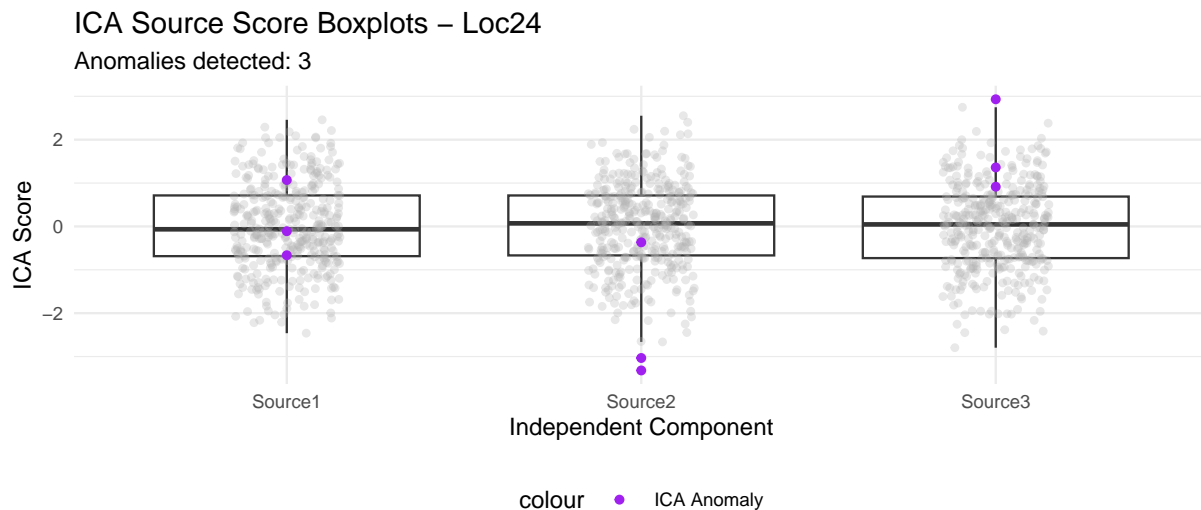

Figure 10: Boxplot for 1 ICA Location with Anomalies

By definition, ICA components often have non-gaussian, heavy-tailed distributions, so the boxplots for ICA components typically show skewness and occasional extremely high or low points like in these last 2 sources. These ICA-specific outliers correspond to short, localized perturbations rather than systematic shifts, explaining why ICA's anomaly detections differ the most from PCA and FA.

The breakdown of anomaly types across all 26 locations reveals a striking pattern: although PCA, FA, and ICA are applied to the same daily traffic curves, the vast majority of anomalies are not jointly detected by all three methods. PCA accounts for the largest portion of anomalies overall, while FA and ICA detect fewer and more selective subsets. This result is fully consistent with the methodological roles of each decomposition. PCA identifies anomalies that manifest as large-scale deviations along the axes of highest variance—for example, abnormally high daily volume, suppressed flow, or significant timing distortions in
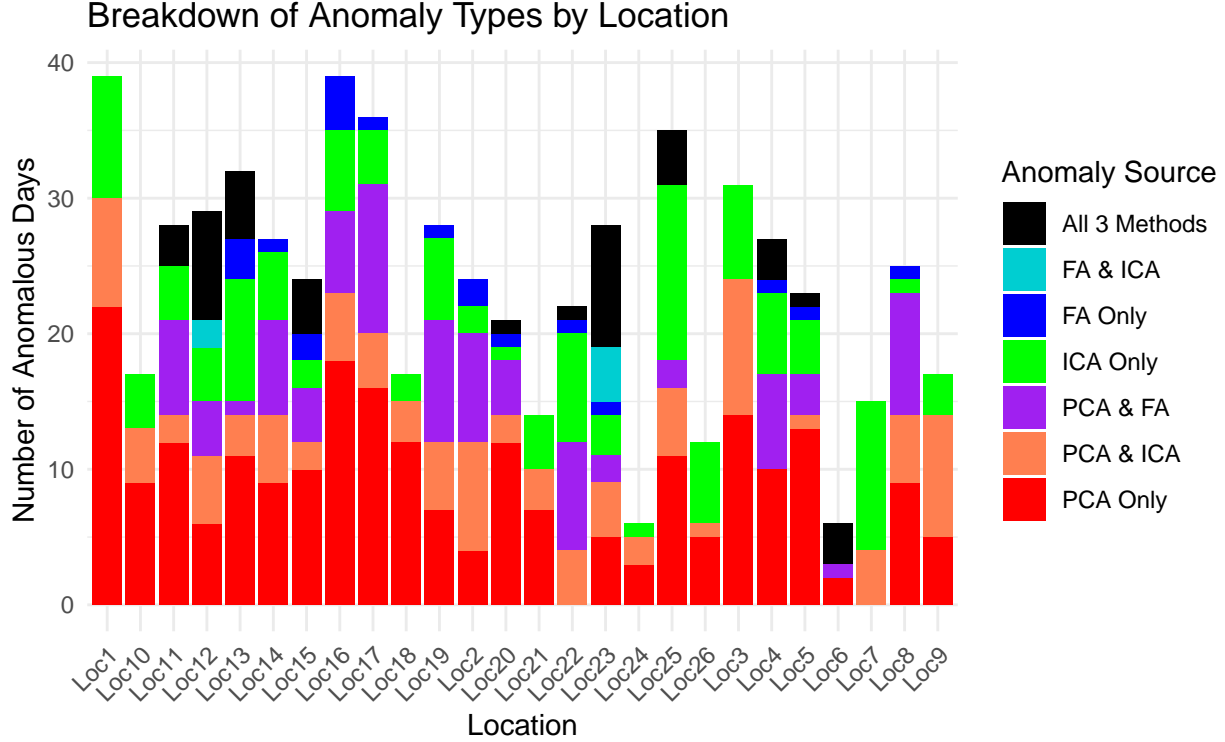
11

Figure 11: Proportion of Total Anomaly Counts

morning or evening peaks. Because daily traffic curves contain substantial natural variability, PCA's variance-maximizing nature makes it highly sensitive to any global shift that pulls a day outside the main data manifold, producing relatively high anomaly counts.

In the context of the project's central goal—identifying and characterizing atypical traffic days at each location—this idiosyncrasy between methods is not a weakness but a strength. The differences in anomaly detection highlight that PCA, FA, and ICA provide complementary anomaly signatures. PCA acts as a broad, variance-aligned detector of major disruptions; FA isolates shape-based irregularities that contradict dominant commuting regimes; and ICA uncovers independent micro-events that reveal fine-grained operational issues. Together, the lack of complete overlap validates the design of the analysis: anomalies in traffic are multi-layered, and understanding them requires a decomposition-based multivariate lens rather than reliance on any single method. If the locations were labeled, there would be further opportunity to reverse engineer anomaly causes based on the contexts of university life and the goals of each multivariate method.

The temporal anomaly sequence plots reveal that anomalies are not uniformly distributed across the 384-day period; instead, they occur in method-specific bursts that reflect distinct types of underlying disruptions. Locations 7 and 23 exhibit concentrated clusters of anomalies around specific day ranges, suggesting periods of sustained irregularity in traffic flow that could correspond to broader system-level disruptions (like construction phases, seasonal events, road closure due to parade). In contrast, other locations show sporadic anomalies, indicating that disruptions at these sites are more idiosyncratic and likely reflect

local micro-events rather than persistent structural changes. Since the dataset is supposed to mimic the traffic congestion at the University of Toronto, something could be said about large groups of students leaving lectures every hour and automatic walk signs. This reinforces the need for a multi-method detection framework, as relying on any single method would obscure the nuanced ways in which traffic deviates over the year.
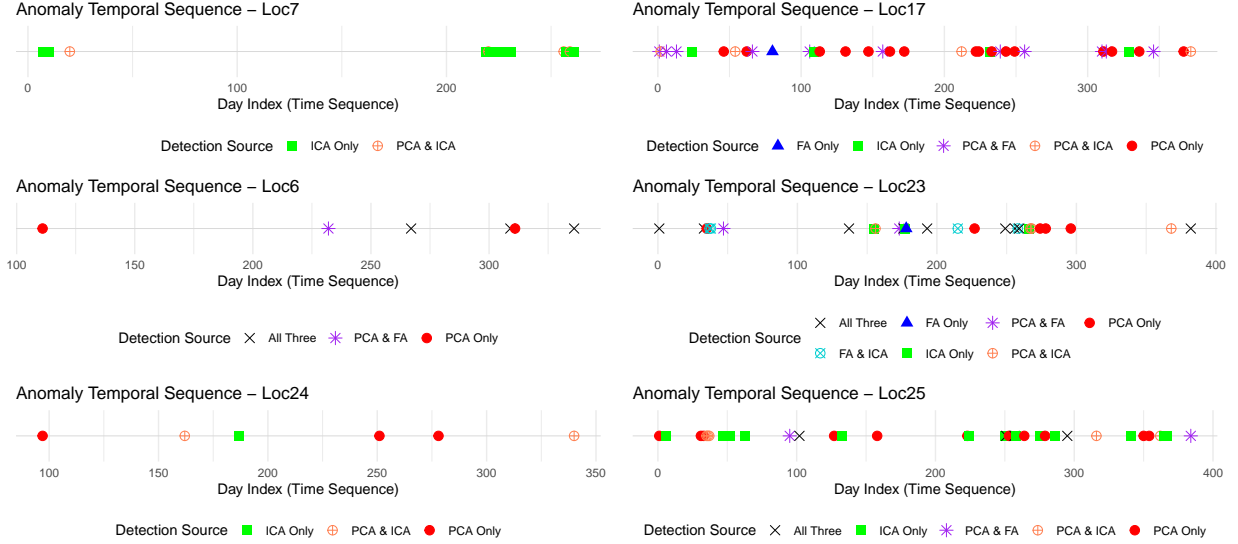

Figure 12: Timeline of Select Anomalies

The PCA-space visualization of clustered anomalies reveals several meaningful and distinct groups, even though the clustering was performed across all locations and based on only the common first 2 PCs. The silhouette analysis selected $K = 5$, indicating that anomalies naturally separate into five coherent behavioral modes rather than forming a cloud. The olive cluster is dense and tightly packed at the center, going by PC1 and PC2 definitions, could mean that these anomalies may be caused by light fluctuations in overall traffic volume rather than structural changes in the daily curve. The remaining clusters were spatially distinct, reflecting more extreme pattern deviations such as shifted peak times, abnormal rush-hour intensities, or overall curve distortions. The presence of similar anomaly types across many locations suggests that some unusual days reflect network-level disruptions, while the smaller clusters indicate rarer, high-severity events. The clear separation between clusters demonstrates that PCA scores effectively encode the structural variability of traffic profiles, enabling interpretable anomaly grouping.

Discussion

A comparative framework utilizing Principal Component Analysis (PCA), Factor Analysis (FA), and Independent Component Analysis (ICA), which provides a multi-faceted approach to traffic anomaly detection, moving beyond single-method reliance, was successfully employed. The criteria for feature retention (90% variance for PCA/FA) and outlier detection,
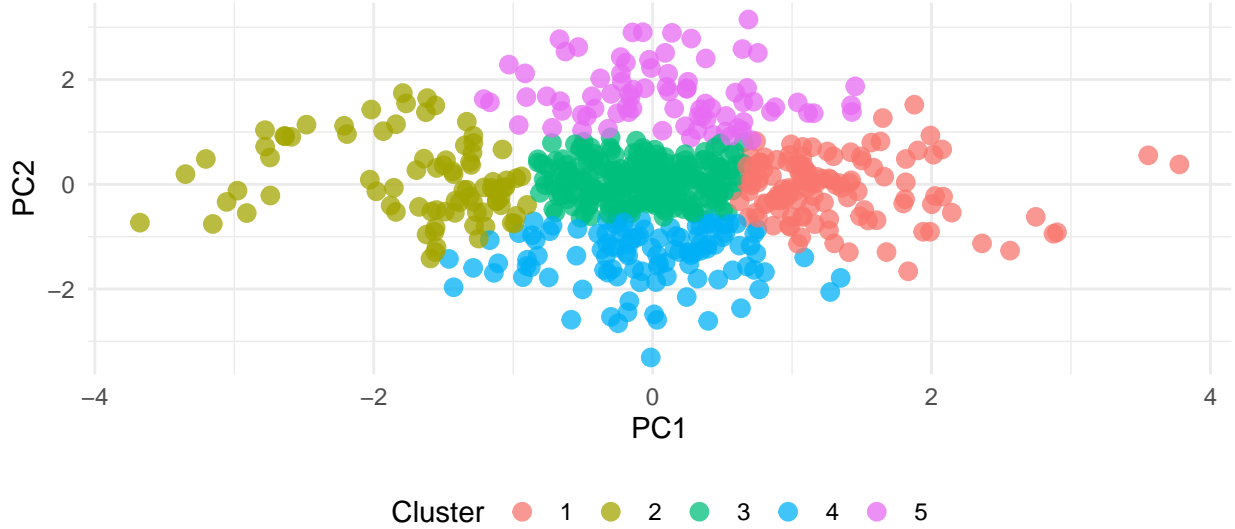
Figure 13: K-means Cluster of Anomalies by 2 PCs

demonstrated that the three methods provide complementary and largely non-overlapping views of the traffic profile, validating the need for a multi-method benchmark. Findings confirm that while PCA serves as a robust dimension reduction strategy, it has fundamental limitations as a standalone detector due to the uninterpretability of its components. Then, with FA and ICA justifying the interpretability of each location's components, the anomalies are explained in a manner that moves beyond simple statistical independence to uncover cause-effect relationships.

There are several avenues for future research. Firstly, all three decomposition methods are fundamentally linear transformations. Although there is beauty in simplicity and it is innocuous to assume that nature always takes the path of least resistance, this assumption may inadequately capture highly non-linear relationships inherent in complex traffic dynamics, such as chaotic system behavior or congestion cascade effects, in which an urban planner should rightfully err on the side of overly pedantic (choose to value false positives) to mitigate risk. Non-linear dimensionality reduction techniques, such as Kernel PCA or manifold learning methods, could potentially reveal subtle, non-linear structural patterns missed by the current approach.

The current reliance on statistical heuristics (like residual thresholds or covariances for PCA/FA) lacks direct calibration against ground-truth external events (confirmed accidents, more features), limiting the assessment of method sensitivity and specificity. While ICA certainly addresses the issue of not having enough information in covariances, providing more factors to ICA would also improve prediction as there have been recent advances in analyzing the reliability of components through statistical testing. Next steps relate to collecting more features of traffic like noting the number of lanes in traffic, speed limits, pedestrian crossings. In addition to upgrading ICA, collecting more features would enable a supervised machine learning validation framework with high performance comparison methods like ensemble learning, which is better suited for big-data problems than ICA is. This would optimize and

expand the geographical capability of traffic-volume anomaly detection. A brief example would be that these would provide helpful insights to commuters planning for a long road trip, which may span various geographical regions in one instance.

The final limitation is that inherently independent per-location analysis is not the detection of system-wide anomalies or coordinated disruptions across multiple sensors. The goal of network-level is to define normalcy, and anomalies relative to the entire system. Then, issues that are invisible or misinterpreted when analyzing individual sensors in isolation can be addressed, as network-level anomaly detection shifts the focus from local fluctuations to the health and behavior of the entire system. Advanced methods like Tensor Decomposition could model the shared temporal and spatial factors simultaneously. By incorporating the network's topological structure, future models can achieve a more precise decoupling of systemic factors from local fluctuations, providing the context, coherence, and causal insight needed for resilient and proactive traffic management across the entire urban network.

Acknowledgements

References

Columbia University Mailman School of Public Health. (n.d.). *Spatiotemporal analysis.*

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13*(4–5), 411–430.

Ma, T., Yao, F., & Zhou, Z. (2024). Network-level traffic flow prediction: Functional time series vs. Functional neural network approach. *The Annals of Applied Statistics*, *18*(1), 424–444.

Piech, C., & Ng, A. (2013). *K-means.*

Ringberg, H., Soule, A., Rexford, J., & Diot, C. (2007). Sensitivity of PCA for traffic anomaly detection. *ACM SIGMETRICS Performance Evaluation Review*, *35*(1), 109–120. https://doi.org/10.1145/1269899.1254895

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Shang, H. L., & Hyndman, R. J. (2010). *Exploratory graphics for functional data.*