Multivariate Routes Through Traffic Anomalies

Erin Xu

December 1st, 2025

Introduction

Due to the unavoidable nature of traffic congestion in urban locations, studying its patterns and underlying dynamics enables daily commuters and transportation authorities to transition from reactive management to proactive intervention, leading 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.).

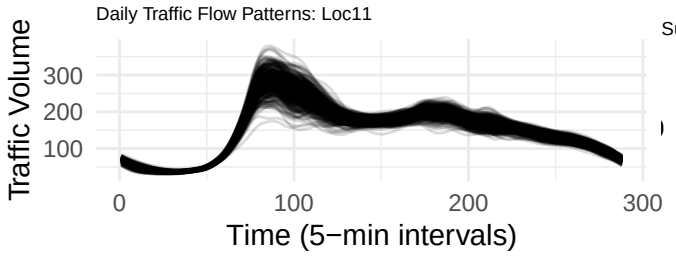
As existing prediction has evolved from interval-based pointwise in univariate time-series data to a functional approach at the network-level utilizing neural networks (Ma et al., 2024), this project offers a comparative assessment of common multivariate statistical techniques that are highly interpretable 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, interpret latent regimes, 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.

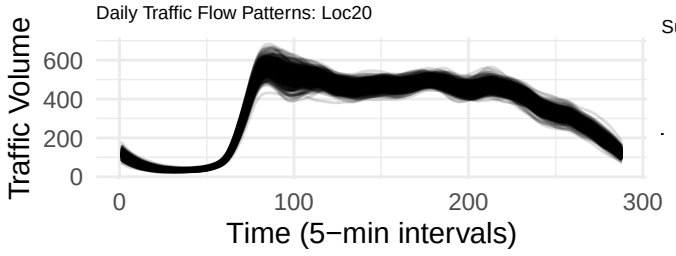
No data cleaning was required, as all locations share uniform dimensions and contain no missing observations. Exploratory visualizations were produced for two randomly selected locations to conserve space. Spaghetti plots by location and by day, accompanied by summary statistics, highlight clear daily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



Summary for Loc11

1st Qu.	Median	Mean	3rd Qu.
94	163	150	192

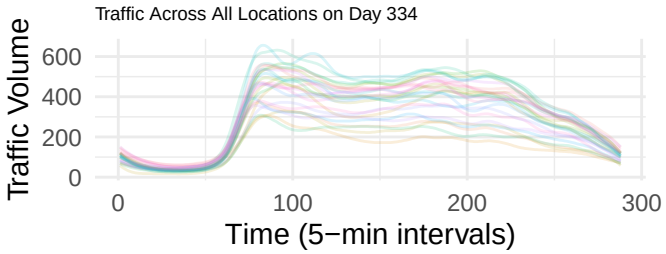


Summary for Loc20

1st Qu.	Median	Mean	3rd Qu.
167	432	341	482

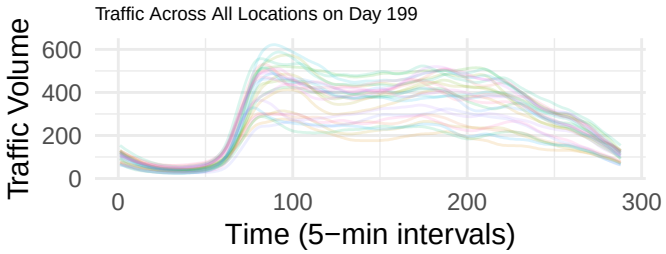
Summary for Day 334

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



Summary for Day 199

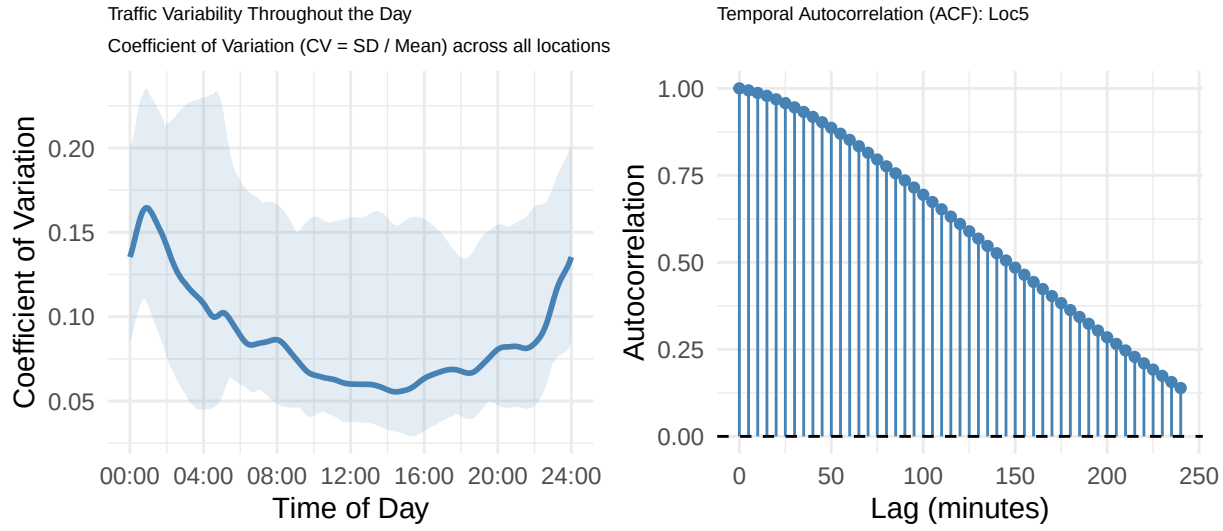
1st Qu.	Median	Mean	3rd Qu.
132.0	274.0	270.6	403.0



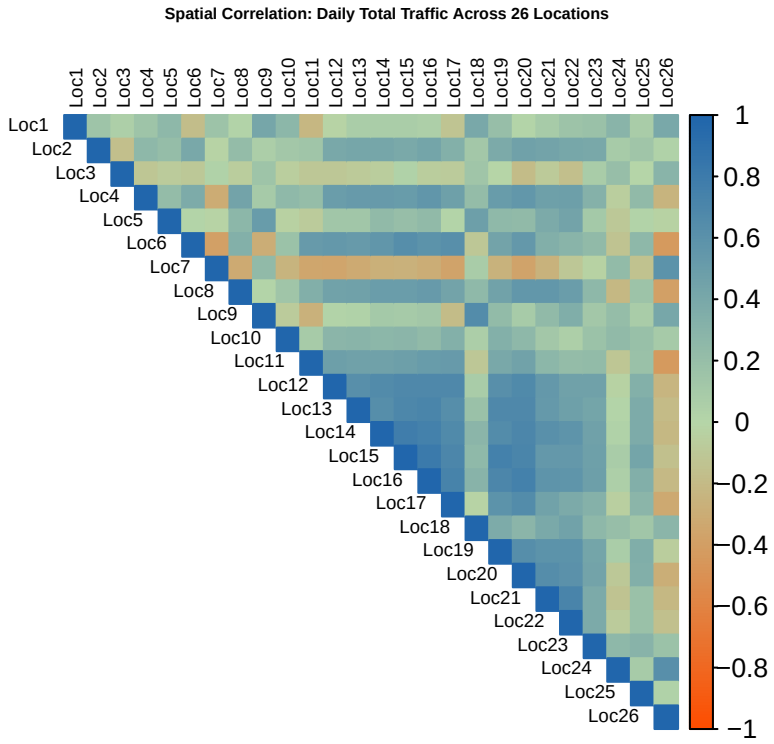
The CV and ACF analyses indicate that the traffic system exhibits a strong diurnal rhythm and persistent temporal dependence, supporting the use of dimension-reduction methods. CV quantifies day-to-day variability at each 5-minute interval. Averaged across locations, the CV curve shows highest variability 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 detect is meaningful with location specific nuances, and network-wide patterns could also be coherent enough as the system has shared temporal dynamics.



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 $V = [v_1, \dots, v_p]$ denote the orthonormal loading functions (the empirical eigenbasis), and let $S = X_cV$ 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 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. 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). 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. FA postulates 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.

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 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. ICA assumes whitened inputs with identity covariance

and therefore requires PCA preprocessing, like in FA (Hyvärinen & Oja, 2000). ICA yields the decomposition $S_{\text{PCA}} = AS_{\text{ICA}}$, where S_{ICA} contains the source signals (scores) and A is the mixing matrix (loadings).

ICA anomalies were detected via the same $1.5 \times \text{IQR}$ 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.

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. 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

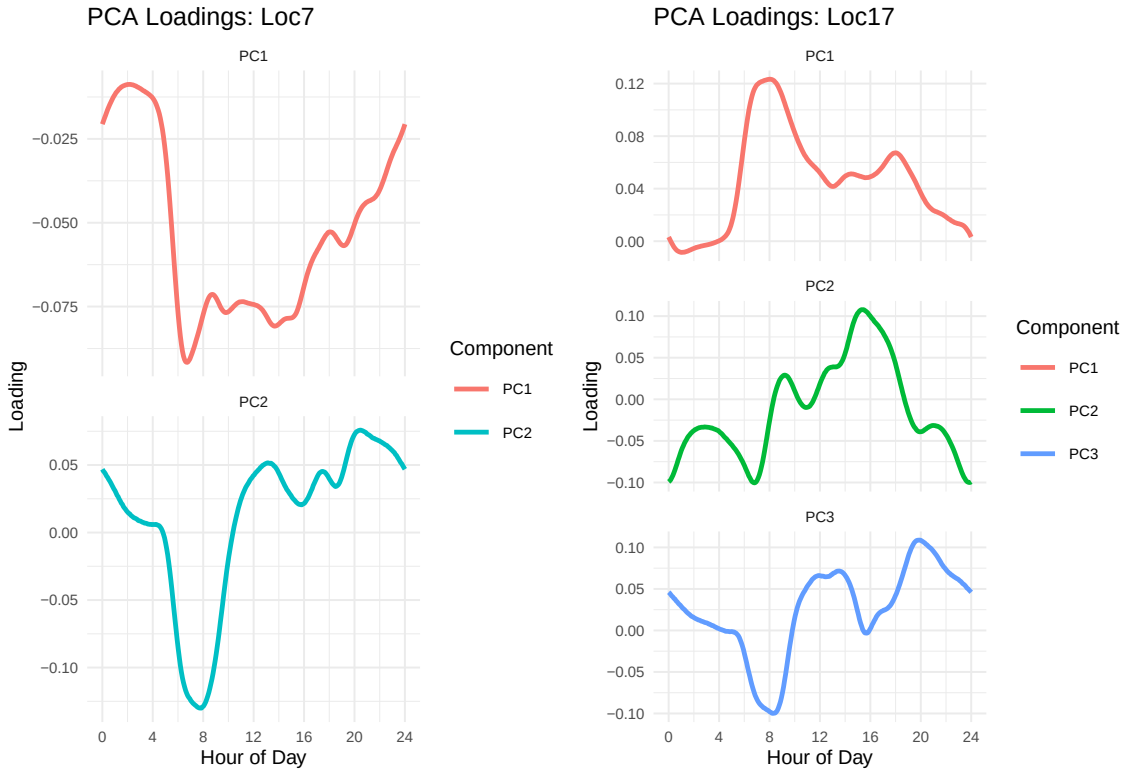
PCA retained between 2 and 9 components across all 26 locations, consistently explaining at least 90% of total variance, with cumulative explained variance ranging from 0.9006 to 0.9404 (Table X). PCA detected between 4 and 31 anomalous days per location, with most locations falling in the 15–25 anomaly 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 / 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.



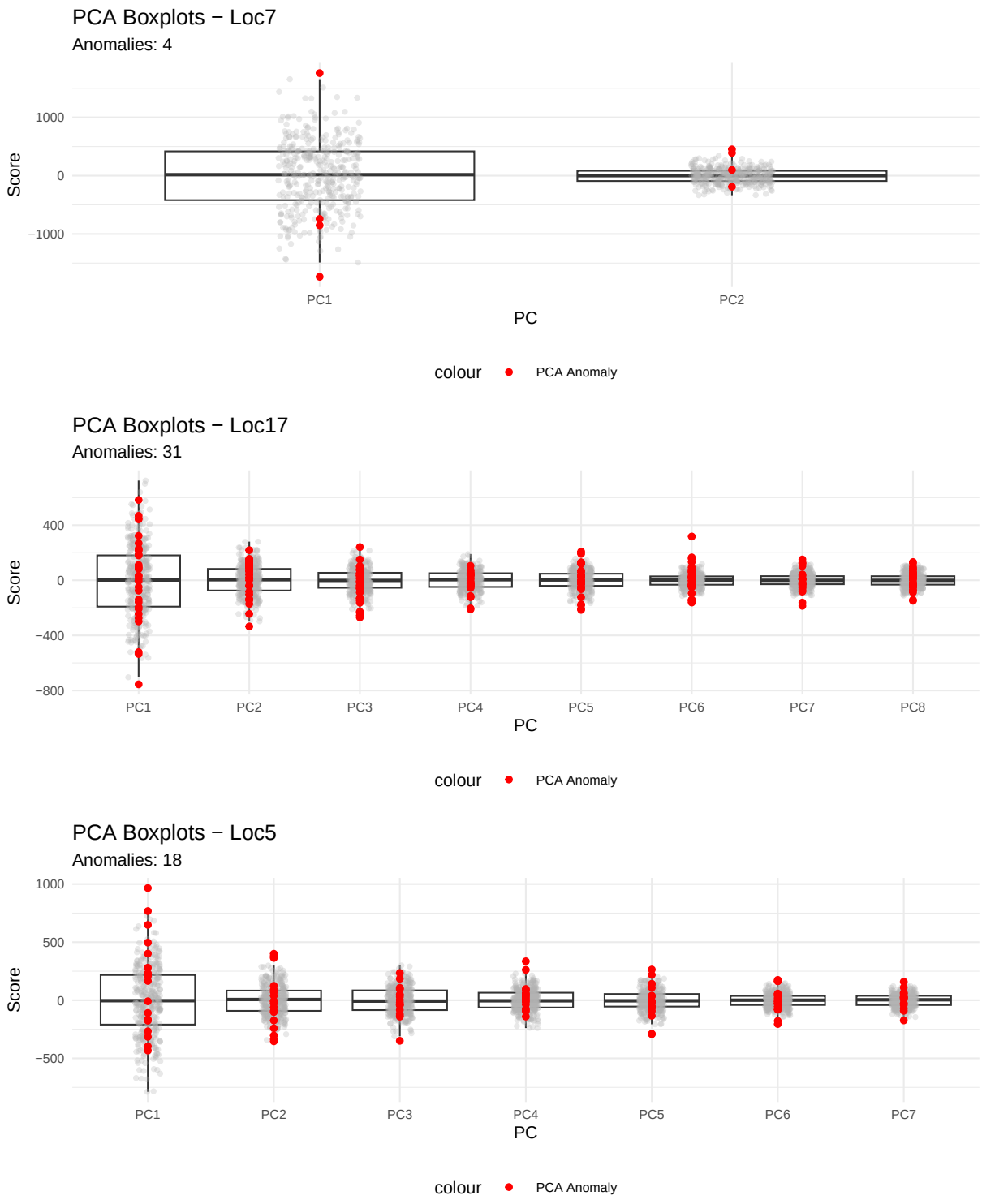
PCA detected:

Minimum anomalies: 4, at Location 7

Maximum anomalies: 31, at location 17

Typical range: 18 anomalies per location

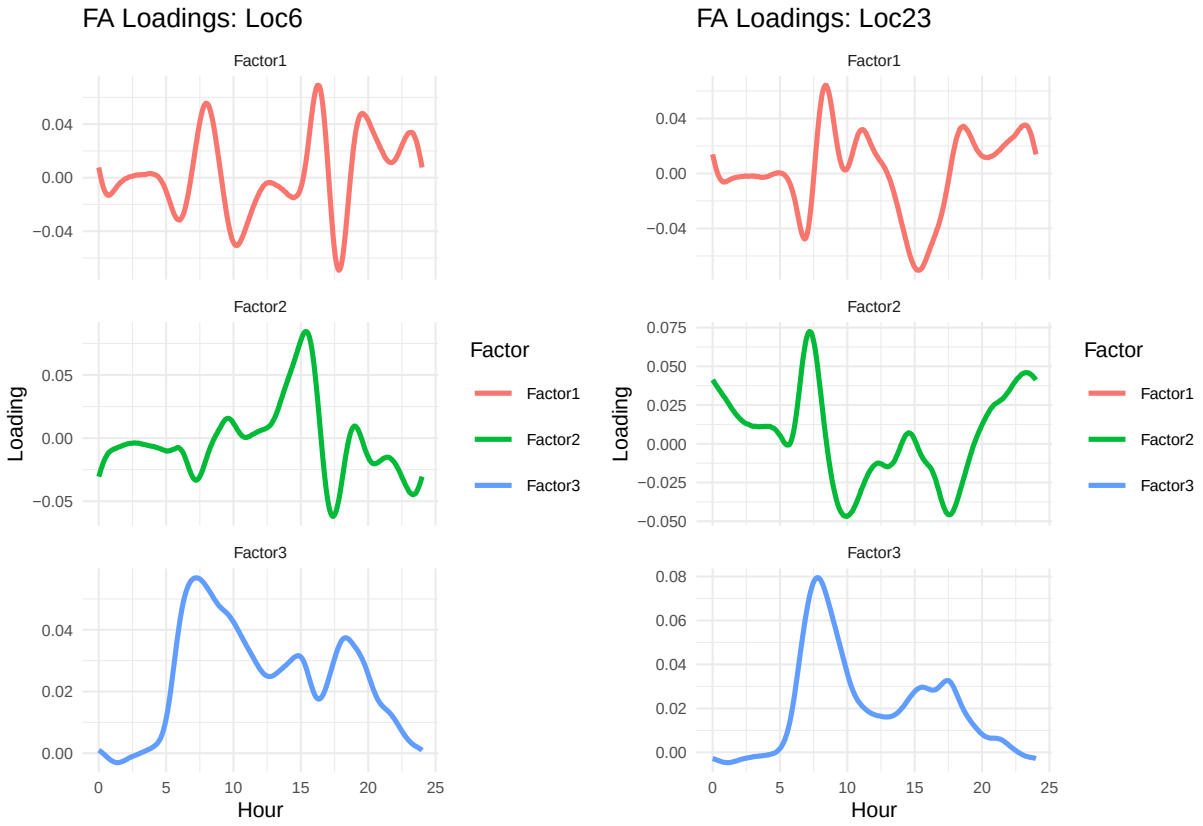
Intuitively, these anomalies correspond to days with unusually large or small PC scores relative to the interquartile range for at least one retained component. Because PCA aligns with variance-maximizing directions, these anomalies could represent large-scale distortions such as full-day surges, holiday 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).

Across locations, the first two FA loadings followed these interpretations:

1. FA1 (baseline daily volume)
2. FA2, FA3 (Shape distortions): They are not variance-maximizing directions.



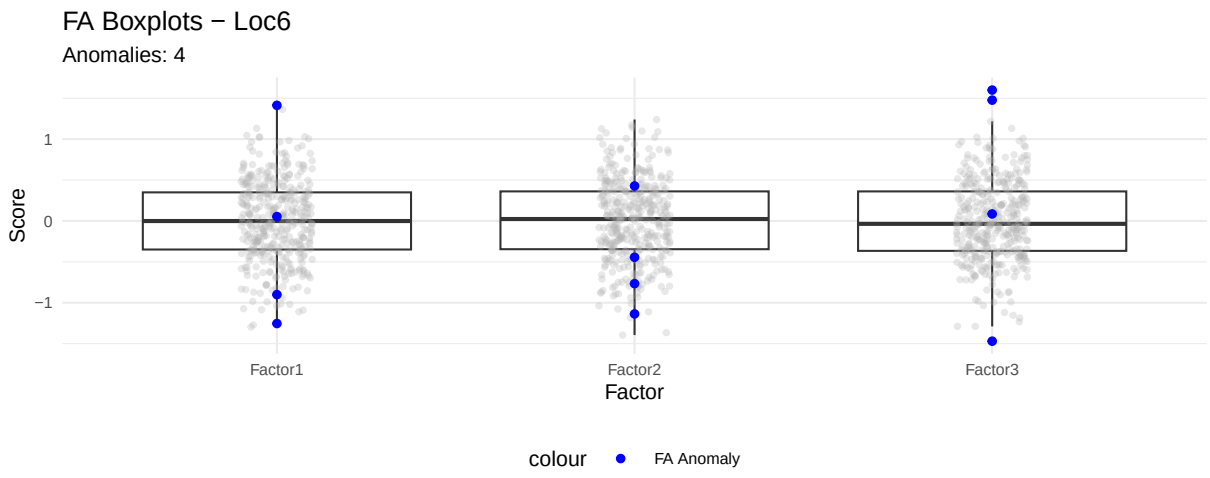
FA detected:

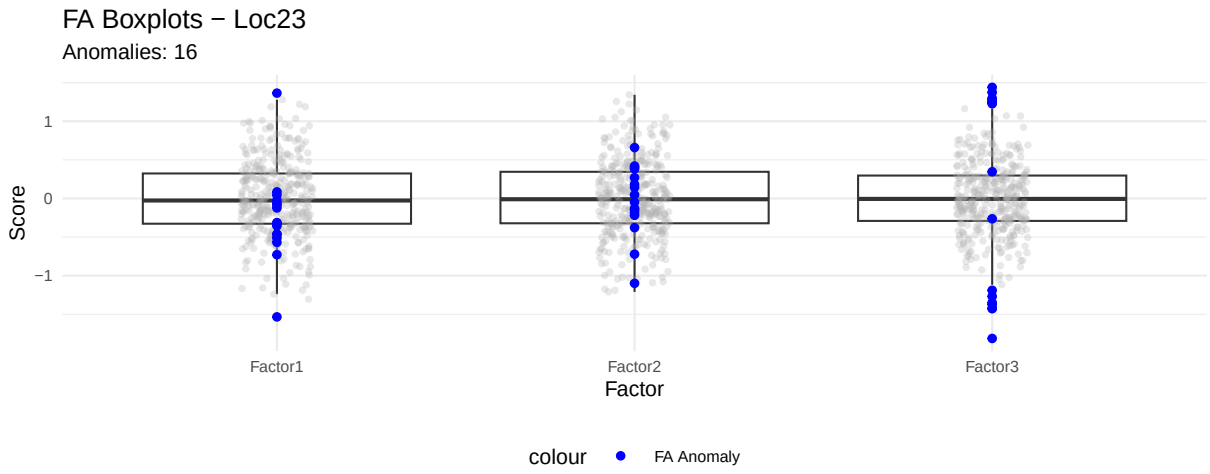
Minimum: 4 anomalies, at Location 6

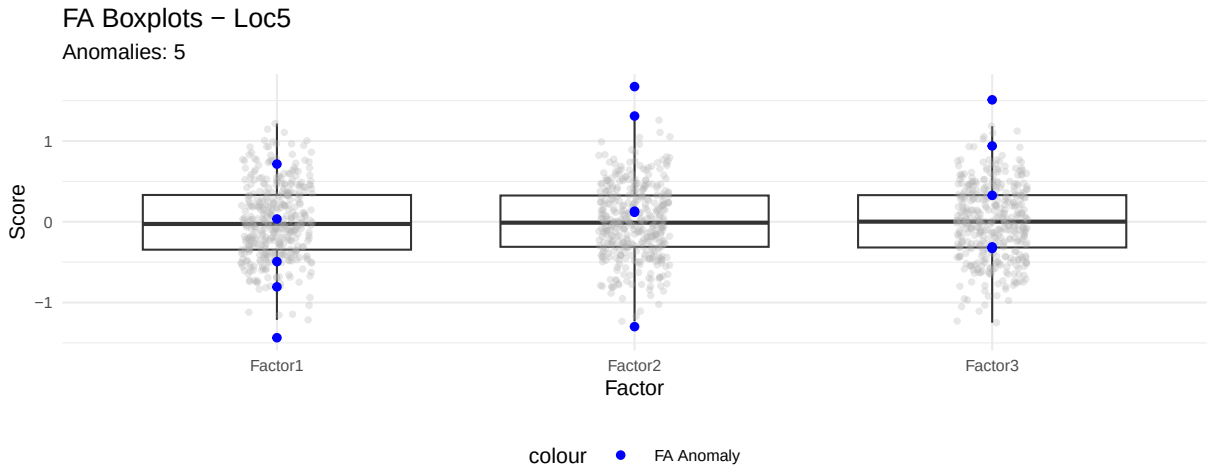
Maximum: 16 anomalies Location 23

Typical: 10 anomalies per successful location

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.







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 two sources followed consistent 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.

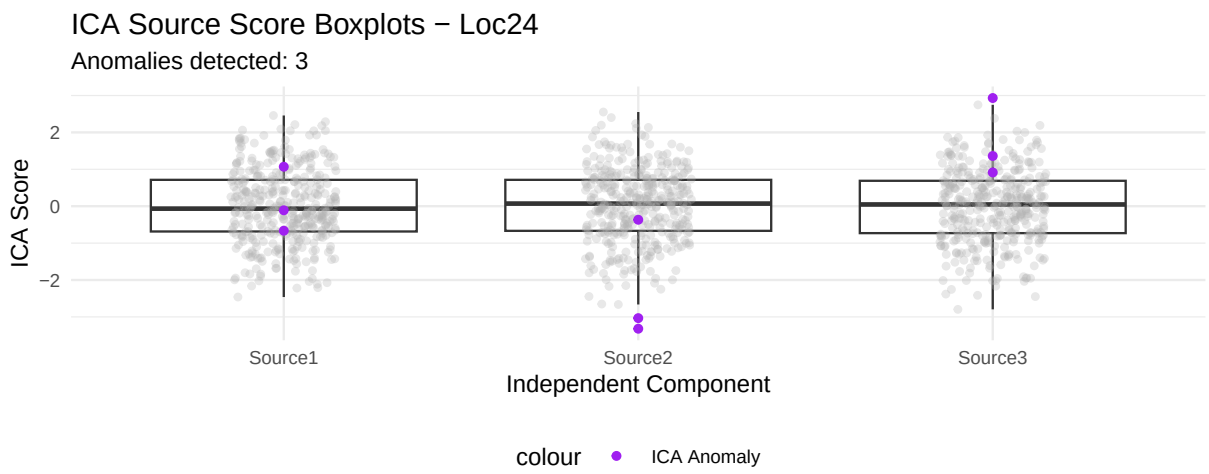
ICA detected:

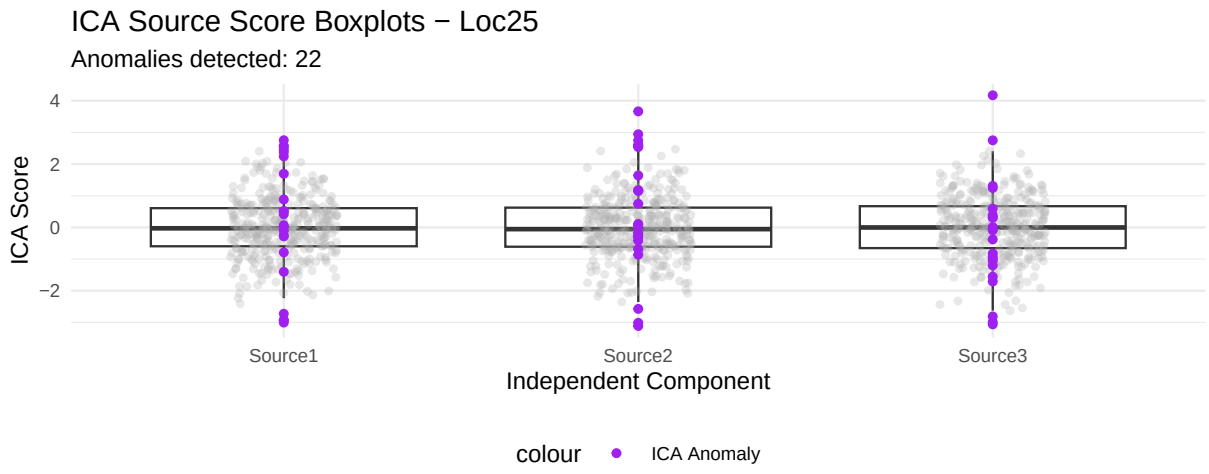
Minimum: 3 anomalies, at Location 24

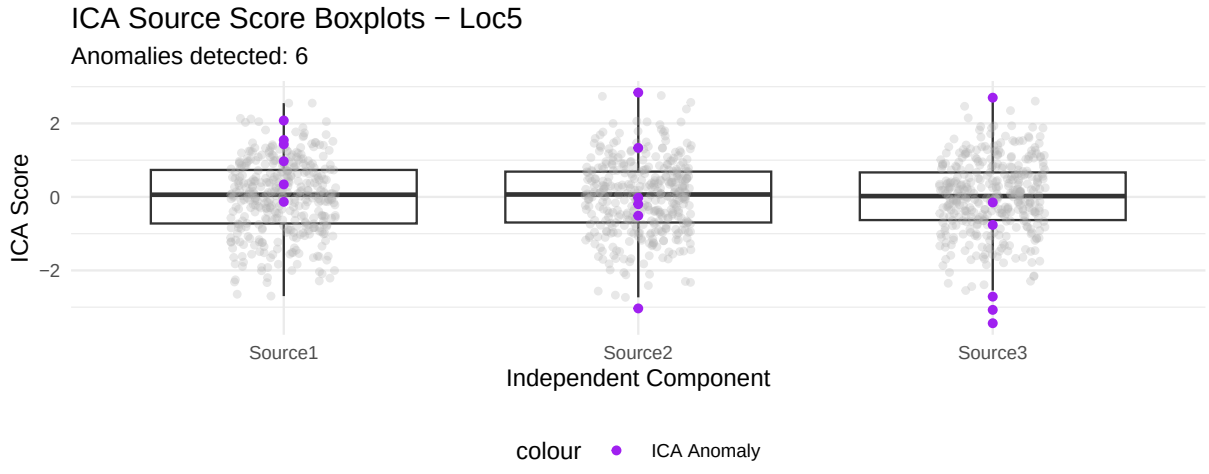
Maximum: 19 anomalies, at Location 25

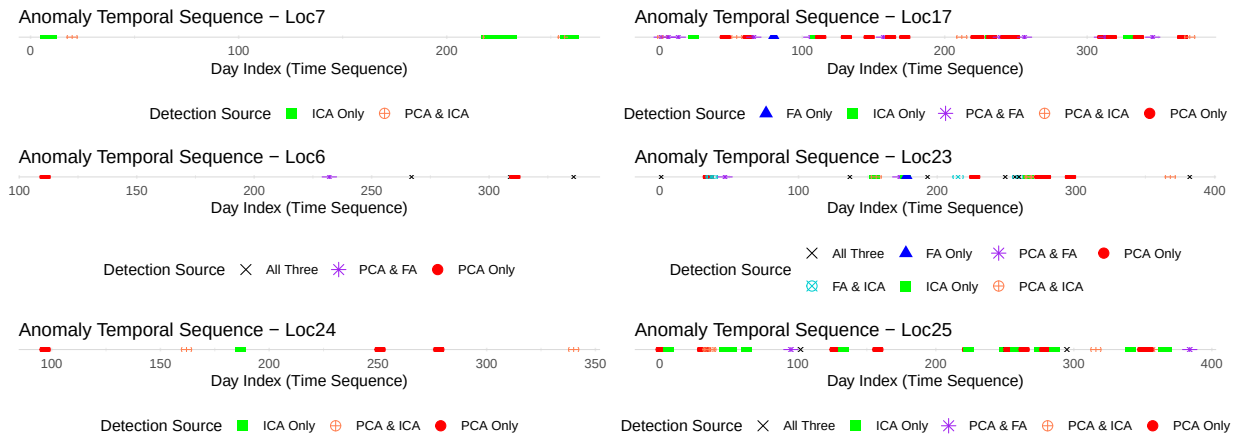
Typical: 10 anomalies per location

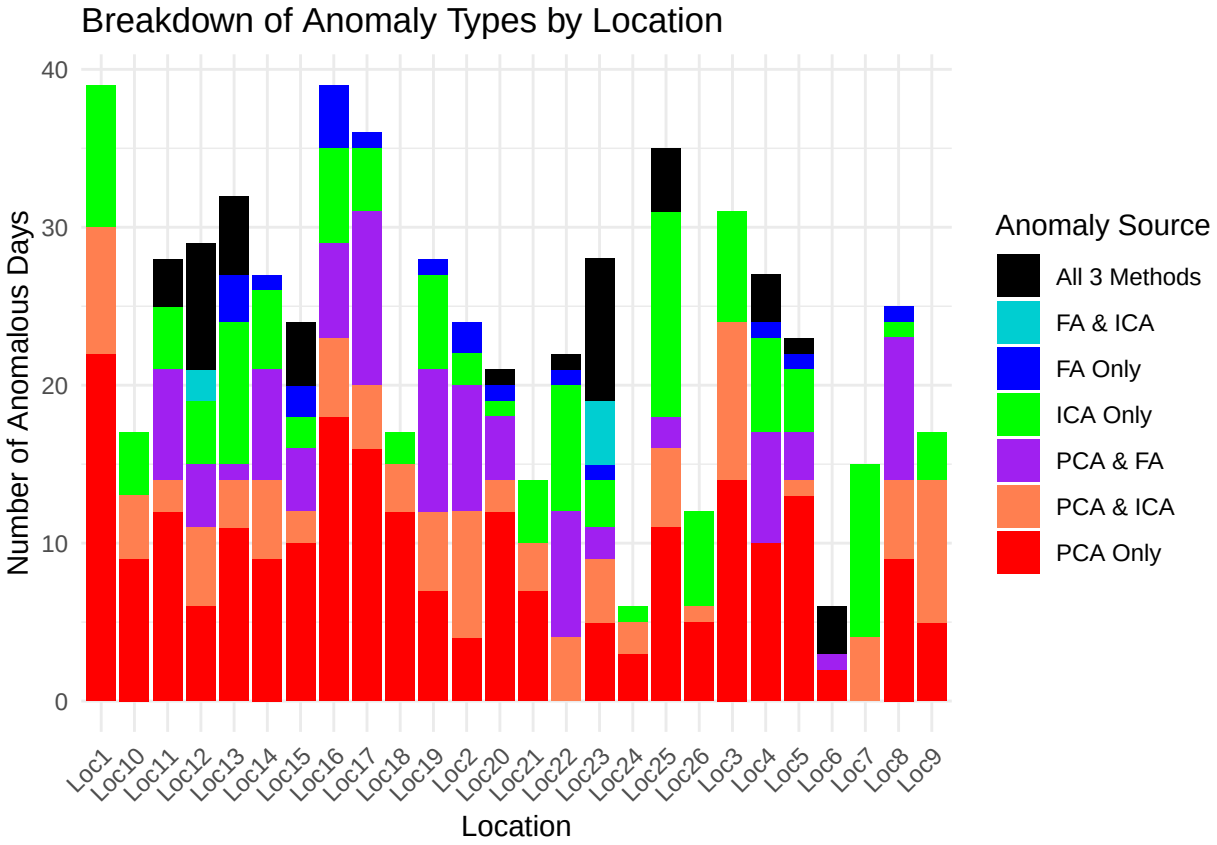
Intuitively, ICA highlights micro-structure anomalies that PCA smooths out and FA diffuses across latent factors. ICA anomalies often overlapped with PCA and FA anomalies, but identified unique, fine-grained disturbances, consistent with ICA's sensitivity to localized, high-frequency deviations. These could be: abrupt, short-lived spikes, dips at specific time intervals, independent transient disruptions (perhaps weather or accidents).


ICA Source Score Boxplots − Loc24
Anomalies detected: 3
ICA Score
2
0
−2
Source1
Source2
Source3
Independent Component
colour
ICA Anomaly


ICA Source Score Boxplots − Loc25
Anomalies detected: 22
ICA Score
4
2
0
−2
Source1
Source2
Source3
Independent Component
colour
ICA Anomaly

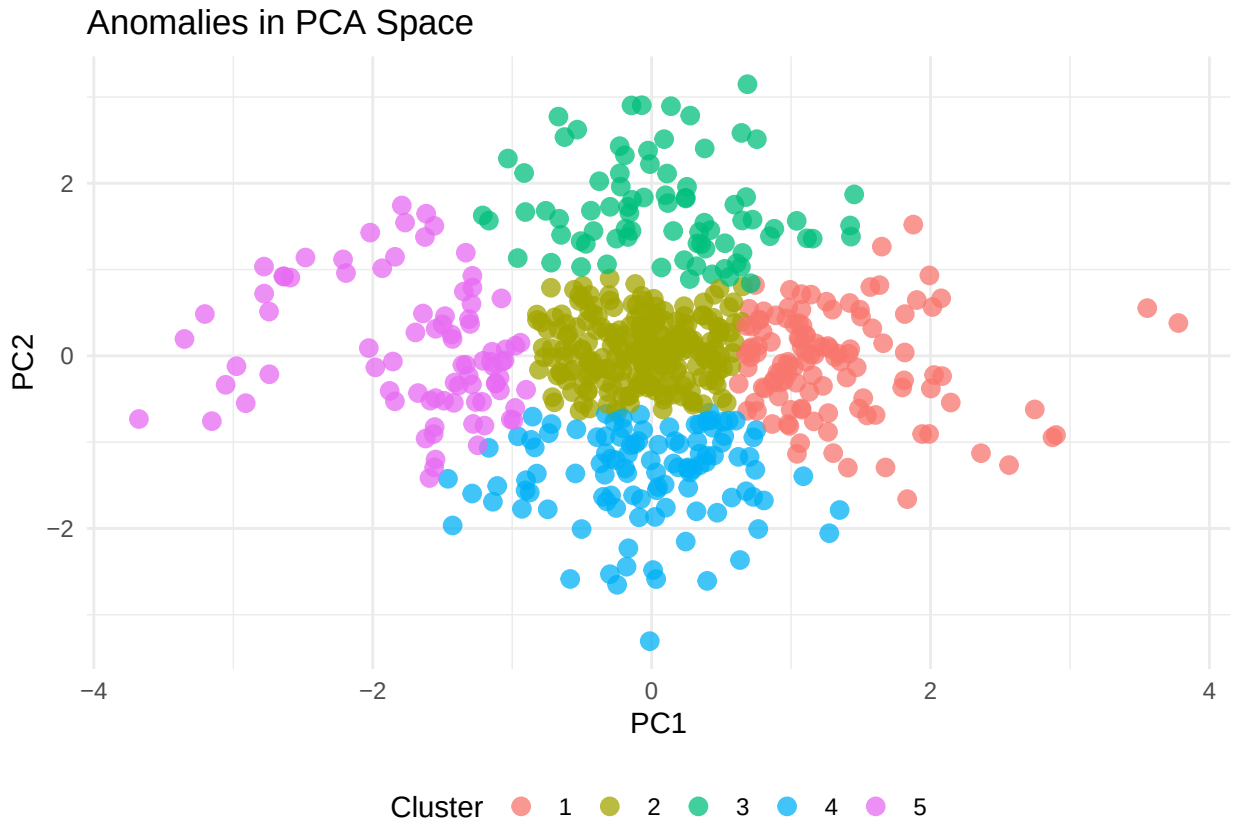

ICA Source Score Boxplots − Loc5
Anomalies detected: 6
ICA Score
2
0
−2
Source1
Source2
Source3
Independent Component
colour
ICA Anomaly





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) to provide a multi-faceted approach to traffic anomaly detection, moving beyond single-method reliance, was successfully employed. Findings confirm that while PCA serves as a robust variance-maximizing benchmark, its fundamental limitations as a standalone detector are frequently understated, and FA and ICA are inherently limited and unrobust alone.

While the joint application of PCA, FA and ICA successfully provided complementary insights into traffic profile anomalies, spanning variance-based deviations, covariance structure shifts, and independent micro-events—several limitations, several avenues for future research are clear. Firstly, all three decomposition methods are fundamentally linear transformations. This assumption may inadequately capture highly non-linear relationships inherent in complex traffic dynamics, such as chaotic system behavior or congestion cascade effects. 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 just statistical heuristics (like residual thresholds for PCA/FA and IQR outliers for ICA sources) lacks direct calibration against ground-truth external events (confirmed accidents, more features), limiting the assessment of method sensitivity and specificity. To overcome this, the project should move towards collecting more features of traffic like noting the number of lanes in traffic, width in roadways, speed limits, pedestrian crossings. This would open the gates for a supervised machine learning validation framework and would allow for the rigorous comparison of method performance using tools like Receiver Operating Characteristic (ROC) analysis and enable the optimization of a data-driven threshold.

The final limitation of the current study is the independent per-location analysis, which 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.

References

The Gemini Flash 2.5 model was used to assist with the formatting of this section.

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*.