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 at the network level (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. There are no missing values and the dimensions are uniform, so there was no data cleaning to be done.

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—ensuring that anomalies detected are robust to modeling assumptions and interpretable in terms of their functional, temporal, and structural characteristics.

Results

Discussion

Traffic dynamics can be intuitively viewed by their diurnal patterns, such as the sharp volume peaks observed during morning and evening rush hours. Conversely, traffic anomalies include deviations from these established norms, such as a sudden isolated car accident, holiday traffic, or systemic events like severe weather closures. Distinguishing these routine fluctuations from true anomalous events requires multivariate methods capable of decomposing the aggregate traffic flow into its underlying normal and abnormal source signals. Different multivariate decompositions reveal

Appendix

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*.