Multivariate Routes Through Traffic Anomalies

Erin Xu

December 1st, 2025

Introduction

Given the unavoidable nature of traffic congestion in urban locations, studying its patterns and underlying dynamics enables the daily commuter as well as transportation authorities to transition from reactive management to proactive intervention, leading to reduced congestion, lower emissions and improved commuter safety around high-density areas. However, accurately detecting and predicting anomalies that cause significant delays remains a challenge due to the inherent complexity of traffic dynamics, which are continuous, stochastic, spatiotemporally autocorrelated and highly cross-correlated at the network level (Columbia University Mailman School of Public Health, n.d.).

As existing studies have evolved from interval-based pointwise prediction of univariate time-series data to a functional approach at network-level prediction with 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 project is to identify and classify location-specific anomalies from the intraday pattern 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, 384 days each, and 288 five-minute time points per day. Then each slice $X \in \mathbb{R}^{\mathbb{n}\times\mathbb{p}}$, 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.

Methodology

The dataset is high-dimensional, functional (recall $p = 288$ time points per day), and correlated, so the ideal methods reduce dimensionality while preserving intepretable time-of-day structure .

PCA is the core method used in this project because each daily traffic profile is a $p = 288$ dimensional vector, and PCA assumes that although each data point x_i lives in a higher dimension $\mathbb{R}^p$, the structure of the data lies on a low dimensional manifold, as such, smooth daily traffic regimes. This is also the main step in functional outlier detection procedure, to express curves in a basis, and PCA is the most computationally efficient. For each location, the 384×288 traffic matrix was transposed so that rows were days and columns were time points. PCA was run on the centered daily traffic matrix, $X_c = X - \mathbf{1}_n \bar{x}^T$. Then, $V = [v_1, ..., v_p]$ is the orthonormal principal directions (loadings, also the empirical basis functions for each location), $X_C V$ contains the associated scores, and the j-th principal score for the day i is $s_{ij} = X_{c,i}^T v_j$. The number of retained components k was chosen by the 90 variance-explained criterion: $\frac{\Sigma_{j=1}^{k} \lambda_j}{\Sigma_{j=1}^{p} \lambda_j} \geq 0.90$, where $\lambda_j = \sigma_j^2$ is the variance explained by the component j, which can be obtained through eigendecomposition of the sample covariance matrix. This balances capturing major cycles and discards higher-frequency noise.

Anomalies were identified directly from the PCA score matrix, $S = [s_{ij}] \in \mathbb{R}^{\mathbb{n} \times \mathbb{k}}$, because PCA scores are uncorrelated. Since outliers often manifest as extreme component-wise deviations in specific principal components, for each component j, the $1.5 \times IQR$ rule was applied to identify outliers. A day i is flagged as an anomaly if $s_{ij} < Q_1 - 1.5IQR$ or $s_{ij} > Q_3 + 1.5IQR$ for any principal component j. In comparison to Mahalanobis distance, the 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. Boxplots give interpretable, location-specific anomaly sets that can be classified by time-of-day (Shang & Hyndman (2010)).

FA is another dimension reduction method, although it assumes that observed random vectors x are generated from latent random vectors called z, then, the covariance amongst variables is due to the influence of latent variables. FA is stricter about the rank of its sample covariance matrix, which needs to be full-rank and invertible, so in practice FA is always done via estimating covariance via iterated PCA until the variances of the error terms converge. Then we obtain latent factor scores and factor loadings.

FA anomalies were also detected using the same $1.5 \times IQR$ rule applied to the factor score matrix F. Different covariance structures are being compared, especially as the original sample covariance matrix is low-rank. FA is not variance maximizing, but instead models latent factors, so intuitively FA anomalies reflect days whose pattern is inconsistent with the latent factors even if their variance is not large (for example, shifted peaks), meanwhile PCA anomalies can be intuitively thought of as being able to capture volume deviations along the main directions where the data usually varies (for example, huge spikes due to an accident).

As a brief extension, ICA was applied to the PCA scores to extract independent signals that have mixed together in the raw time-series (Hyvärinen & Oja, 2000). ICA is unstable when applied to highly correlated features as well, because it assumes an uncorrelated covariance matrix (I), so it requires inputs be whitened, dimension-reduced and variance-normalized, which is exactly what PCA does. The scores, called source signals S, loadings, called the mixing matrix A, were obtained.

ICA anomalies were detected with the same $1.5 \times IQR$ rule applied to the source scores

S. 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.

Following their indentification through the methods above, the anomalies were clustered in their 2D principal component space via k-means clustering, due to the range of principal components of the anomalies. K-means was also selected because ofits suitability for continuous feature space as it partitions by Euclidean distance as well as its computational efficiency as an unsupervised learning algorithm (Piech & Ng, 2013). The number of clusters k was determined by calculating the average silhouette coefficient after computed on each anomaly i out of N anomalies, $k = \arg\max_k \frac{1}{N}\Sigma_{i=1}^{N}\frac{b(i)-a(i)}{max(a(i),b(i))}$, the inner functions are the average distance from one observation to all others in its assigned cluster and the minimum average distance to observations in any other cluster (Rousseeuw, 1987).

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.*