

Statistical Analysis and Predictive Modelling of Urban Traffic Flow for Route Generation: Evidence from Padua and Port Harcourt

Fun-Akpo Pere Kormane¹; Thankgod Ode²; Captain Gospel Otto^{3*}; Kemejika Ichechi Amadi-Oparaeli⁴

^{1,2,3,4}Department of Civil Engineering, Rivers State University, Port Harcourt, Nigeria

Corresponding Authors: Captain Gospel Otto^{3*}

³ORCID ID: 0000-0001-6855-6871

Publication Date: 2026/10/12

Abstract: Urban route generation depends on traffic statistics that are persistent in time, uneven in space and only partly predictable. This study applies descriptive, temporal, concentration, clustering, regression and sensitivity statistics to two complementary datasets: 7,908 ten-minute junction-volume periods and hourly travel times from Padua, Italy, and aggregate morning and evening peak observations from seven Port Harcourt, Nigeria, locations linked to a directed OpenStreetMap network of 48,504 nodes and 111,839 edges. Padua network flow showed near-complete short-term persistence (lag-1 autocorrelation = 0.994) and strong weekly recurrence (lag-7-day autocorrelation = 0.981); weekday flow exceeded weekend flow by 24.49%. In Port Harcourt, total peak flow rose by only 5.04% from morning to evening, yet the between-location coefficient of variation roughly doubled (0.180 to 0.372) and the normalised Herfindahl concentration rose from 0.0046 to 0.0198. A paired comparison of the seven locations found no statistically detectable morning-to-evening difference ($t(6) = 0.48$, $p = 0.647$), and flow was not correlated with reported speed ($r = -0.10$ and 0.09), showing that loading alone does not identify congestion. K-Means clustering supported two traffic regimes (silhouette = 0.465; Davies-Bouldin = 0.863). On chronologically held-out data, Gradient Boosting gave the lowest error for traffic volume (RMSE = 12.704 vehicles; $R^2 = 0.977$) and for travel time (RMSE = 218.274 s; $R^2 = 0.673$), with current and lagged observations dominating permutation importance. Route-overlap statistics showed that 41 of 42 directed origin-destination pairs had a single sufficiently distinct route, and demand scaling to 1.5 times base raised total travel time by about 54%. The results show that statistical evidence can quantify both what is predictable and what the network structure prevents routing from changing.

Keywords: Traffic Flow Statistics; Autocorrelation; Coefficient of Variation; Herfindahl Index; K-Means Clustering; Gradient Boosting; Route Generation; Port Harcourt.

How to Cite: Fun-Akpo Pere Kormane; Thankgod Ode; Captain Gospel Otto; Kemejika Ichechi Amadi-Oparaeli (2026) Statistical Analysis and Predictive Modelling of Urban Traffic Flow for Route Generation: Evidence from Padua and Port Harcourt. *International Journal of Innovative Science and Research Technology*, 11(10), 292-300. <https://doi.org/10.38124/ijisrt/26oct253>

I. INTRODUCTION

Recurrent congestion reduces mobility, raises operating costs and weakens road safety, and it is most acute where urban growth has outpaced network expansion (Faheem et al., 2024; Hammoumi et al., 2025; Mystakidis et al., 2025). Improving how demand is distributed over an existing network is a practical alternative to continuous road building, but any route-generation method is only as sound as the statistical description of traffic on which it rests. Route costs, bottleneck rankings and demand allocations all inherit the persistence, dispersion and noise of the underlying counts and travel times.

Traffic-flow theory supplies the deterministic relationships among flow, speed and density (Greenshields, 1935; Helbing, 1997; Garber & Hoel, 2019), whereas recent machine-learning studies emphasise forecasting accuracy (Abduljabbar et al., 2025; Emenike et al., 2026; Nofal, 2024; Zhu et al., 2025). Less attention has been given to the intermediate statistical questions that determine whether a forecast can be trusted for routing: how persistent the series is, how unevenly demand is spread across locations, whether apparent differences are distinguishable from sampling variation, how many distinct operating regimes exist, how much of the variance a model leaves unexplained, and how

sensitive downstream allocations are to forecast error and demand growth.

This article addresses those questions using data assembled for an optimised route-generation framework for urban traffic flow. Its contributions are as follows:

- It quantifies temporal dependence, daily and weekly recurrence and between-junction dispersion in a high-resolution traffic series from Padua, Italy.
- It characterises spatial concentration of morning and evening peak flows at seven Port Harcourt, Nigeria, locations using dispersion, concentration and rank-based inferential statistics, and tests whether flow alone explains congestion.
- It selects traffic regimes with internal validity indices and compares Linear Regression, Random Forest and Gradient Boosting on chronologically separated training, validation and test data, with permutation importance.
- It reports network-structure and scenario statistics (route overlap, edge-loading dispersion, demand scaling and systematic prediction-error perturbation) that bound what statistical prediction can deliver in routing practice.

The remainder of the article reviews related work, describes the data and statistical methods, presents results in four groups, and discusses implications and limitations.

II. RELATED WORK

- Statistical description of traffic. Empirical traffic studies conventionally begin with volume, speed and travel-time summaries, peak-hour profiles and measures of variability. Field studies in African cities have used such descriptive evidence to characterise congestion at intersections and along corridors, including Addis Ababa (Sokido, 2024), Enugu (Chukwurah et al., 2025) and the Lokoja–Benin corridor (Murisela et al., 2025). Congested travel in large urban areas is typically concentrated in a minority of road segments and time periods (Çolak et al., 2016), which motivates dispersion and concentration measures in addition to means.

- Forecasting and learning methods. Systematic reviews describe a broad movement from classical time-series models to ensemble and deep learning for traffic flow, speed and congestion prediction (Alomari et al., 2022; Attiou & Lahby, 2025; Xing et al., 2023; Yao et al., 2023). Applied studies in developing and emerging contexts include congestion prediction in Casablanca (Hammoumi et al., 2025), traffic-condition prediction on Ikorodu Road, Lagos (Imoh & Rad, 2025), and accident-aware safe routing in Ethiopia (Berhanu et al., 2024). Tree-ensemble methods such as Random Forest (Breiman, 2001) and gradient boosting (Friedman, 2001) are widely used because they capture nonlinear interactions with modest tuning, while linear regression remains an essential baseline (Bishop, 2006; Hastie et al., 2009).
- Unsupervised structure and limited data. Clustering has been used to identify operating states and to support transfer when local data are scarce (Zou & Chung, 2024). Internal validity indices such as the silhouette coefficient (Rousseeuw, 1987) and the Davies–Bouldin index (Davies & Bouldin, 1979) allow the number of clusters to be chosen from the data rather than imposed.
- Gap addressed here. Evaluation in this literature is dominated by accuracy metrics (Abduljabbar et al., 2025; Emenike et al., 2026; Zhu et al., 2025). Fewer studies connect temporal dependence, spatial concentration, inferential caution on small local samples, regime structure, model error and scenario sensitivity within one analysis aimed at route generation. This article does so for a data-constrained setting in which dense records exist for one city and only aggregate observations exist for the target city.

III. DATA AND STATISTICAL METHODS

➤ Data Sources and Treatment

Two datasets played distinct roles and were never treated as interchangeable. The Padua data supplied temporal depth for statistical characterisation and model development; the Port Harcourt data supplied locally observed magnitudes for calibration and engineering interpretation. Table 1 summarises them.

Table 1 Datasets Used in the Statistical Analysis

Dataset	Coverage	Verified size	Use
Padua junction volumes	40 junctions; 6 Feb–1 Apr 2026; 10-min	7,908 periods (6,755 with at least one valid count)	Temporal statistics, regimes, volume prediction
Padua hourly travel times	Directed junction pairs; hourly	500,721 rows (432,803 non-null means)	Travel-time prediction
Padua aggregate travel times	Directed movements	1,538 records	Descriptive support
Port Harcourt peak flows	7 locations; morning and evening peaks; four-week survey	14 location-period observations	Calibration, concentration and inference
OpenStreetMap network	Port Harcourt, motor-vehicle roads	48,504 nodes; 111,839 directed edges	Connectivity, route redundancy

The 1,153 Padua periods missing at all 40 junctions were treated as unavailable rather than zero. Six valid periods with a network total of zero were retained because they were recorded values. Timestamps were converted to Europe/Rome local time, and an hourly total was formed only when all six ten-minute totals were present. The published travel-time files contain only origin, destination and mean travel time (plus a time bin in the hourly file); no median or observation-count fields were available, so none is analysed.

➤ Descriptive, Temporal and Concentration Statistics

For each junction, the mean, median, standard deviation, maximum and coefficient of variation (CV = standard deviation / mean) were computed from valid counts. Temporal dependence in the network total $Q(t)$ was measured by the Pearson correlation between $Q(t)$ and $Q(t - k)$ at lags from ten minutes to seven days, using only timestamp pairs for which both values existed, so that missing blocks were never bridged.

$$r_k = \text{corr}(Q_t, Q_{t-k})$$

Spatial concentration of peak flow across n locations with shares s_i was measured by the Herfindahl–Hirschman index and its normalised form, which equals zero when all locations are equally loaded (Hirschman, 1964).

$$H = \sum_{i=1}^n s_i^2, \quad H^* = \frac{H - 1/n}{1 - 1/n}$$

Table 2 Chronological Sample Sizes

Task	Train	Validation	Test
Traffic volume (records)	221,280	14,240	34,120
Travel time (records)	140,177	9,637	20,623

Linear Regression served as the baseline. Random Forest used 160 trees, maximum depth 18 and minimum leaf size 2. Histogram-based Gradient Boosting used a learning rate of 0.06, up to 300 iterations, 31 leaf nodes, L2 regularisation of 0.10 and early stopping after 25 non-improving iterations (an extreme-gradient-boosting

Because only seven locations were observed, inferential statements about them are deliberately limited. For this article, the morning and evening flows were additionally compared with a paired t-test and a Wilcoxon signed-rank test (Wilcoxon, 1945), and associations between flow, reported speed and the congestion index (a combination of normalised loading and relative speed deficit) were examined with Pearson and Spearman correlations. These tests were computed from the tabulated aggregate observations and are descriptive of this sample, not population-level evidence.

➤ Regime Identification

K-Means clustering (Lloyd, 1982) was applied to the training period using standardised network flow, mean junction flow, cross-junction variability and ten-minute change, for $K = 2$ to 6. The number of regimes was chosen using the silhouette coefficient (higher is better), the Davies–Bouldin index (lower is better), inertia and the size of the smallest cluster.

➤ Predictive Models and Evaluation

Two supervised tasks were defined: next ten-minute junction volume (269,640 junction-time records) and next-hour directed-pair mean travel time (170,437 records). Predictors were current conditions, three consecutive lags, cyclical hour and weekday terms, location identity and traffic regime. Records adjacent to missing blocks were removed using exact timestamp checks. Data were split chronologically, not randomly, into training, validation and test sets (Table 2), so that test performance reflects genuinely future periods.

implementation was replaced because its runtime library was unavailable in the execution environment). Models were selected on validation error and reported once on the untouched test set using mean absolute error, root mean square error and the coefficient of determination.

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|, \quad \text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}, \quad R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Permutation importance, the increase in RMSE when one predictor is randomly shuffled, quantified each input's contribution.

➤ Scenario Statistics

Predicted conditions were combined with equal weights of 0.20 for travel time, distance, congestion, relative loading and structural risk to produce dynamic link costs. Candidate routes between the 42 directed origin–destination (OD) pairs of the seven locations were screened with a 0.80 overlap rule. Robustness was examined by scaling model-derived OD demand to 0.8, 1.0, 1.2 and 1.5 times base demand and by

applying systematic travel-time perturbations from –20% to +20% in seven steps; dispersion across scenarios was summarised by the CV.

IV. RESULTS I: DESCRIPTIVE AND TEMPORAL STATISTICS

➤ Junction-Level Distributions

Ten-minute counts were highly dispersed at every major junction (Table 3). Junction agg_11 carried the greatest mean volume (220.05 vehicles per period), followed by agg_6 (176.43) and agg_9 (148.25), with CVs between 0.686

and 0.712. Junction agg_35 had a mean of only 7.73 vehicles but the largest relative variability (CV = 1.614) and a median (2.00) far below its mean, indicating a right-skewed,

intermittent series. High CV therefore reflects intermittency as much as importance: it should not be used alone to rank sites.

Table 3 Selected Padua Junction Statistics (Ten-Minute Counts, Valid n = 6,755 Each)

Junction	Mean	Median	SD	Maximum	CV
agg_11	220.05	243.00	152.87	565	0.695
agg_6	176.43	207.00	121.12	440	0.686
agg_9	148.25	159.00	105.56	378	0.712
agg_35	7.73	2.00	12.47	84	1.614

For the three busiest junctions the median exceeds the mean, a signature of a bimodal series in which long low-flow stretches coexist with a sustained daytime plateau; this is consistent with the two regimes identified in Section 6.1.

➤ Daily and Weekly Structure

Mean network flow showed a pronounced daily cycle with a secondary morning rise and a stronger late-afternoon peak. The highest hourly mean occurred at 17:00 (32,955.32 vehicles per hour across the observed junctions), followed by 18:00 (32,487.74) and 16:00 (31,240.60); the lowest occurred at 03:00 (897.60). The peak-to-trough ratio is therefore about 37 to 1, so a single daily mean would describe almost no actual hour. Dispersion across dates widened during active hours, meaning that peak demand was both higher and less stable.

➤ Autocorrelation

Network flow was close to deterministic over short horizons and strongly recurrent over longer ones (Table 4). The lag-1 correlation of 0.994 implies that the previous ten-minute value alone accounts for about 99% of linear variance ($r^2 \approx 0.988$). Correlation declined smoothly with lag, reached 0.916 at 24 hours, fell to 0.857 at 48 hours, and recovered to 0.981 at seven days. The dip at 48 hours and recovery at seven days is the expected signature of a weekly cycle in which the same weekday, rather than any day, is the best analogue.

Table 4 Autocorrelation of Padua Network Flow

Lag	Correlation (r)
10 minutes	0.994
20 minutes	0.985
30 minutes	0.973
24 hours	0.916
48 hours	0.857
7 days	0.981

Two consequences follow for modelling. First, any forecaster must be judged against a persistence baseline that is already very strong, which explains why a linear model performs well in Section 6. Second, the high autocorrelation means that successive observations are not independent, so conventional standard errors computed as if they were independent would be too small; chronological validation is therefore essential.

V. RESULTS II: SPATIAL CONCENTRATION AND INFERENCE AT PORT HARCOURT

➤ Peak Flows and Congestion Index

Total recorded peak flow across the seven Port Harcourt locations was 4,271 vehicles per hour in the morning and 4,486.27 in the evening, a net increase of 5.04%. Location-level changes were far larger and of opposite sign (Table 5): Mile 3 Diobu rose 40.68% and Choba 37.60%, whereas Borikiri, Woji, Iwofe and Abuloma fell by 19.74%, 19.70%, 17.31% and 8.83%.

Table 5 Port Harcourt Peak Flows, Reported Speed and Congestion Index (CI)

Location	AM flow (veh/h)	PM flow (veh/h)	Change (%)	Speed (km/h)	AM CI	PM CI
Abuloma	589	537.00	-8.83	49.03	0.410	0.270
Artillery	719	780.00	+8.48	33.78	0.656	0.547
Borikiri	618	496.00	-19.74	3.98	0.889	0.708
Choba	617	849.00	+37.60	23.98	0.685	0.682
Iwofe	387	320.00	-17.31	35.10	0.411	0.303
Mile 3 Diobu	708	996.00	+40.68	35.39	0.631	0.639
Woji	633	508.27	-19.70	27.21	0.663	0.478

The Borikiri speed (3.98 km/h) was retained as reported in the source but flagged, because it is exceptionally low and conflicts with the source discussion. Conclusions below that depend on it are labelled accordingly.

➤ Dispersion and Concentration

Although the network total changed little, spatial dispersion changed substantially (Table 6). The between-

location CV of flow rose from 0.180 in the morning (mean 610.1, SD 109.8 veh/h) to 0.372 in the evening (mean 640.9, SD 238.4 veh/h). The Herfindahl index rose from 0.1468 to 0.1598 against a minimum of $1/7 = 0.1429$, so normalised concentration rose from 0.0046 to 0.0198. Neither period was dominated by one location, but evening demand was about four times more concentrated in normalised terms, driven by the Mile 3 Diobu and Choba surges.

Table 6 Spatial Dispersion and Concentration of Peak Flow (n = 7 Locations)

Statistic	Morning	Evening
Total flow (veh/h)	4,271.00	4,486.27
Mean across locations (veh/h)	610.1	640.9
SD across locations (veh/h)	109.8	238.4
Coefficient of variation	0.180	0.372
Herfindahl index H	0.1468	0.1598
Normalised concentration H*	0.0046	0.0198

➤ Inferential Tests on a Small Sample

The paired morning-to-evening difference averaged +30.75 veh/h (SD = 169.07). A paired t-test gave $t(6) = 0.48$, $p = .647$ (95% CI for the mean difference -125.6 to 187.1 veh/h; Cohen's $d_z = 0.18$), and the Wilcoxon signed-rank test agreed ($W = 13$, $p = .938$). The sample therefore provides no evidence that the typical location carries a different load in the evening. This is itself informative: the evening effect is not a uniform shift but a redistribution, visible in the change in dispersion rather than in the mean. With only seven pairs the tests have little power to detect moderate differences, so the result must be read as absence of detectable evidence, not evidence of absence.

Flow was not associated with speed or with congestion status in this sample. Pearson correlations between flow and reported speed were $r = -.10$ ($p = .831$) in the morning and $r = .09$ ($p = .854$) in the evening; with Spearman's rank correlation for the evening, $\rho = .21$ ($p = .645$). Excluding the flagged Borikiri speed, the evening correlation was $r = -.24$ ($p = .643$). Rank correlations between flow and the CI were

likewise weak (morning $\rho = .32$, $p = .482$; evening $\rho = .29$, $p = .535$). Borikiri illustrates the point: its moderate flow (618 and 496 veh/h) combined with the lowest reported speed gave the highest CI in both periods (0.889 and 0.708). Congestion is thus a joint property of loading and speed deficit, and ranking sites by flow alone would misidentify priorities. These tests are exploratory, computed on aggregate data from one survey, and should be confirmed with link-level measurements.

VI. RESULTS III: REGIME IDENTIFICATION AND PREDICTIVE MODEL COMPARISON

➤ Traffic Regimes

$K = 2$ gave both the highest silhouette coefficient (0.465) and the lowest Davies–Bouldin index (0.863) among the five solutions (Table 7). Inertia fell steadily as K increased, as it must, but separation quality deteriorated, indicating that additional clusters subdivided existing operating states rather than revealing new ones. The smallest cluster shrank from 2,463 periods at $K = 2$ to 258 at $K = 6$.

Table 7 K-Means Diagnostics for Traffic Regimes

K	Inertia	Silhouette	Davies–Bouldin	Smallest cluster
2	11,379.3	0.465	0.863	2,463
3	8,607.6	0.389	0.949	1,423
4	6,460.3	0.363	0.925	689
5	5,439.0	0.351	0.954	568
6	4,652.5	0.338	0.958	258

The two regimes differ by a factor of about five in network flow (Table 8). The low-flow regime averaged 928.74 vehicles per ten-minute period, against 4,823.96 in the moderate-flow regime; the label is relative to the Padua sample and does not imply absence of congestion. Cross-junction variability was higher in the low-flow regime (CV =

0.855 versus 0.706), meaning that quiet periods were spatially uneven. The regimes recovered the daily cycle without peak hours being specified: low flow dominated from midnight to the early morning and after about 21:00, and the higher regime dominated from the morning build-up through the evening peak.

Table 8 Characteristics of the Two Selected Regimes

Regime	Network flow	Mean junction flow	Junction CV	Ten-minute change
Low-flow	928.74	23.22	0.855	-6.00
Moderate-flow	4,823.96	120.60	0.706	5.29

➤ Predictive Performance

Gradient Boosting had the lowest validation RMSE on both tasks (Table 9). For traffic volume its validation RMSE of 13.816 was 11.8% lower than the linear baseline (15.668) and 1.5% lower than Random Forest (14.033). For travel time, the corresponding reductions were 4.3% (213.718

versus 223.310 s) and 2.1% (versus 218.289 s). On the untouched test set, Gradient Boosting achieved RMSE = 12.704 vehicles, MAE = 8.278 and $R^2 = 0.977$ for volume, and RMSE = 218.274 s, MAE = 102.460 s and $R^2 = 0.673$ for travel time.

Table 9 Training, Validation and Test Performance

Task	Model	Set	MAE	RMSE	R^2
Volume	Linear Regression	Train	9.740	14.936	0.968
Volume	Linear Regression	Validation	10.260	15.668	0.967
Volume	Random Forest	Train	5.465	8.282	0.990
Volume	Random Forest	Validation	9.207	14.033	0.974
Volume	Gradient Boosting	Train	8.335	12.697	0.977
Volume	Gradient Boosting	Validation	9.025	13.816	0.975
Volume	Gradient Boosting	Test	8.278	12.704	0.977
Travel time	Linear Regression	Train	104.661	213.282	0.650
Travel time	Linear Regression	Validation	116.358	223.310	0.623
Travel time	Random Forest	Train	60.480	122.615	0.884
Travel time	Random Forest	Validation	106.975	218.289	0.640
Travel time	Gradient Boosting	Train	91.275	191.922	0.717
Travel time	Gradient Boosting	Validation	105.537	213.718	0.655
Travel time	Gradient Boosting	Test	102.460	218.274	0.673

Three patterns matter statistically. First, the strong linear baseline (volume $R^2 = 0.967$ on validation) is a direct consequence of the autocorrelation in Section 4.3: most of the signal is persistence. Second, Random Forest fitted the training data most tightly but generalised worst relative to its training fit (travel-time R^2 of 0.884 in training against 0.640 in validation), whereas Gradient Boosting's volume RMSE was almost identical in training (12.697) and test (12.704), showing little overfitting under early stopping. Third, travel time remained harder to predict: test $R^2 = 0.673$ leaves roughly one-third of variance unexplained, and the residual distribution had a longer upper tail, so unusually long travel times remain the least predictable cases. Gradient Boosting

also trained faster (about 1.65 s for volume and 0.99 s for travel time) than Random Forest (12.32 s and 18.73 s).

➤ Permutation Importance

Current and lagged observations dominated both models (Table 10). For volume, shuffling the current value raised RMSE by 72.807, far more than any other input; the first and third lags followed (9.117 and 5.033). For travel time, the current value and the first three lags raised RMSE by 21.8 to 40.2 s each, and the cyclical hour term by 8.0 s. The traffic-regime indicator contributed little once current and lagged values were included, so it is more useful for interpretation and transfer than for short-horizon accuracy.

Table 10 Leading Permutation-Importance Results (Increase in RMSE)

Task	Feature	Increase in RMSE
Volume	current	72.807
Volume	lag 1	9.117
Volume	lag 3	5.033
Volume	hour (cosine)	4.245
Volume	lag 2	3.554
Travel time	current	40.175
Travel time	lag 1	38.914
Travel time	lag 3	25.305
Travel time	lag 2	21.787
Travel time	hour (cosine)	8.030

VII. RESULTS IV: NETWORK STATISTICS AND SCENARIO SENSITIVITY

All Port Harcourt route flows, capacities and link loads in this section are model-derived analytical estimates, not observed link-level measurements.

➤ Network and Route-Redundancy Statistics

The extracted network had a beta index of 2.306 and a mean total node degree of 4.575 (median 6). Degree distribution was bimodal: 51.59% of nodes had degree six and 32.57% degree two, and only 5.36% had degree eight or more. All 48,504 nodes belonged to one weakly connected component, and the largest strongly connected component held 48,431 nodes (99.85%). Residential roads made up 65.80% of directed length, while primary, secondary and

trunk roads together made up 5.83%. Links above the 99th percentile of normalised edge betweenness (threshold 0.284) formed continuous bands, indicating that cross-city movement is channelled through a limited backbone.

The key statistic for routing is route independence. Of 42 directed OD pairs, 41 (97.62%) retained a single route after the 0.80 overlap screen and one (Iwofe–Artillery) retained two, giving 43 accepted routes. Accepted routes ranged from 3.42 to 34.34 km (median 13.76; mean 14.91), and the maximum retained overlap was 0.766. Model-derived edge loading was more uneven in the evening (CV = 0.670) than in the morning (0.611) even though mean relative loading was lower (0.239 versus 0.295), mirroring the spatial concentration in Section 5.2.

Table 11 Network Travel Time by Routing Strategy (Vehicle-Hours)

Strategy	Morning	Evening
B0 calibrated baseline	1,546.84	2,816.98
B1 distance-only	1,546.97	2,817.14
B2 current travel time	1,546.71	2,816.81
B3 ML minimum cost	1,546.71	2,816.81
Proposed allocation	1,546.76	2,816.87

The period effect was far larger than the strategy effect. Under the proposed framework, mean travel time rose from 21.73 to 37.67 minutes, mean delay from 8.67 to 24.22 minutes, mean speed fell from 41.08 to 24.89 km/h and maximum utilisation rose from 1.034 to 1.251.

➤ Demand Scaling and Prediction-Error Perturbation

Scaling demand changed outcomes much more than changing strategy (Table 12). At 1.5 times base demand, total

➤ Comparison of Routing Strategies

Because only one OD pair could be redistributed, five routing strategies produced almost identical network statistics (Table 11). In the morning, total travel time ranged only from 1,546.71 to 1,546.97 vehicle-hours (a spread of 0.017%), and in the evening from 2,816.81 to 2,817.14 (0.012%). The proposed allocation improved total time over the calibrated baseline by 0.0052% (morning) and 0.0037% (evening) and sent 80.44% of the Iwofe–Artillery demand to the lower-cost route. These differences are real but operationally negligible, and a significance test would add nothing: the constraint is topological, not statistical.

travel time was 54.21% above base in the morning and 54.12% in the evening, while mean delay rose only 7.0% and 4.3%. Total time therefore scaled roughly in proportion to demand, whereas the count of critical links grew sharply and nonlinearly (morning 6 to 59; evening 31 to 203), identifying the 1.2 to 1.5 range as the onset of widespread overload.

Table 12 Demand Sensitivity of the Proposed Framework

Demand	Period	Total time (veh-h)	Mean delay (min)	Max utilisation	Critical links
0.8 Q ₀	Morning	1,232.36	8.58	0.827	0
1.0 Q ₀	Morning	1,546.76	8.67	1.034	6
1.2 Q ₀	Morning	1,869.88	8.83	1.240	6
1.5 Q ₀	Morning	2,385.28	9.28	1.551	59
0.8 Q ₀	Evening	2,244.49	24.07	1.001	1
1.0 Q ₀	Evening	2,816.87	24.22	1.251	31
1.2 Q ₀	Evening	3,404.81	24.50	1.501	44
1.5 Q ₀	Evening	4,341.49	25.26	1.876	203

Systematic travel-time errors of –20% and +20% changed total time almost exactly in proportion (morning 1,237.41 and 1,856.11 vehicle-hours; evening 2,253.50 and 3,380.25), giving a CV of 0.132 across the seven scenarios in both periods. The Iwofe–Artillery route-1 share stayed at 0.196 in every scenario. Allocation was therefore rank-robust but magnitude-sensitive: a common proportional error preserves the ordering of route costs but scales every predicted burden.

➤ Link Degradation and Closure

Halving the capacity of the selected critical link doubled maximum utilisation (1.251 to 2.501) yet raised total network travel time by only 0.100%, an example of severe local impact hidden in a network average. Full closure forced five

Choba-bound movements onto routes outside the screened set; modelled total time then fell by 2.494% and maximum utilisation fell to 0.943. This counter-intuitive result indicates that routes excluded by cost or overlap rules can become useful under disruption; it is not a recommendation to close links, and the emergency paths require field verification.

VIII. DISCUSSION

- What is predictable. Traffic volume in the Padua series is predictable largely because it is persistent. The lag-1 correlation of 0.994 and the 0.981 seven-day correlation imply that recent values and the same weekday carry most of the information, and the near-parity of Linear Regression ($R^2 = 0.967$) and Gradient Boosting ($R^2 =$

0.975) on validation confirms this. Nonlinear learning added a measurable but modest 11.8% reduction in RMSE. Similar reliance on recent and temporal information is reported in other forecasting studies (Emenike et al., 2026; Nofal, 2024; Zhu et al., 2025). Travel time is less predictable (test $R^2 = 0.673$), and its heavy residual tail means that exactly the long delays that matter most for routing are those the model anticipates least well (Alomari et al., 2022).

- What averages hide. Two results show that central tendency misleads. In Padua, a 37-to-1 ratio between peak and overnight hours makes a daily mean uninformative. In Port Harcourt, a 5.04% rise in total flow concealed a doubling of between-location dispersion (CV 0.180 to 0.372) and an almost fourfold rise in normalised concentration, while the paired test found no detectable mean shift ($p = .647$). Concentration of congested travel in a minority of locations and periods is a recognised feature of urban networks (Çolak et al., 2016), and it argues for dispersion and concentration statistics alongside means in any monitoring programme.
- Flow is not congestion. Flow and reported speed were uncorrelated in the seven-location sample, and the highest congestion index belonged to a site with only moderate flow. Statistical screening based on volume alone would therefore misidentify priority sites. Because the Borikiri speed is flagged, this finding is best taken as a prompt for re-measurement rather than a precise estimate.
- Statistics versus structure. The most important statistical finding for route generation is a limit rather than a score. Strategy differences of about 0.01% in total travel time arise because 41 of 42 OD pairs have one sufficiently distinct route. Better forecasts cannot create independent alternatives; they can only rank those that exist. This qualifies accuracy-centred evaluation in the routing literature (Abduljabbar et al., 2025; Hammoumi et al., 2025) and supports reporting route-redundancy statistics next to error metrics. Where redundancy is absent, the value of prediction lies in diagnosis and scenario testing: demand scaling exposed the rapid growth of critical links between 1.2 and 1.5 times base demand, and the $\pm 20\%$ error test showed that route rankings, though not travel-time magnitudes, are insensitive to proportional bias.

IX. CONCLUSION

Statistical analysis clarifies both the strengths and the limits of data-driven route generation. Padua traffic is highly persistent (lag-1 $r = 0.994$), strongly weekly (lag-7-day $r = 0.981$), and separates into two regimes (silhouette = 0.465). Gradient Boosting predicts next-interval volume with RMSE = 12.704 vehicles and $R^2 = 0.977$ and next-hour travel time with RMSE = 218.274 s and $R^2 = 0.673$, with current and lagged observations dominating importance. In Port Harcourt, a 5.04% rise in total peak flow concealed redistribution (CV 0.180 to 0.372), flow did not track speed, and route-overlap statistics showed that 41 of 42 OD pairs lacked an independent alternative, so that routing strategies differed by about 0.01% in network travel time while a 50% demand increase raised it by about 54%. Future work should add multi-season local counts, link-level validation of model-

derived flows, formal tests of forecast differences and link-specific error scenarios.

REFERENCES

- [1]. Abduljabbar, R., Dia, H., & Liyanage, S. (2025). Machine learning traffic flow prediction models for smart and sustainable traffic management. *Infrastructures*, 10(7), 155. <https://doi.org/10.3390/infrastructures10070155>
- [2]. Alomari, A. H., Khedaywi, T. S., Marian, A. R. O., & Jada, A. A. (2022). Traffic speed prediction techniques in urban environments. *Heliyon*, 8(12), Article e11847.
- [3]. Attioui, M., & Lahby, M. (2025). A systematic literature review of traffic congestion forecasting: From machine learning techniques to large language models. *Vehicles*, 7(4), 142. <https://doi.org/10.3390/vehicles7040142>
- [4]. Berhanu, Y., Schröder, D., Wodajo, B. T., & Alemayehu, E. (2024). Machine learning for predictions of road traffic accidents and spatial network analysis for safe routing on accident- and congestion-prone road networks. *Results in Engineering*, 23, Article 102737. <https://doi.org/10.1016/j.rineng.2024.102737>
- [5]. Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- [6]. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- [7]. Chukwurah, G. O., Okeke, F. O., Isimah, M. O., Nnaemeka-Okeke, R., Okonta, E. D., Awe, F. C., Idoko, A. E., Guo, S., & Okeke, C. A. (2025). Analysis of route-way dynamics in urban traffic congestion of Enugu, Nigeria. *Future Transportation*, 5(2), 71. <https://doi.org/10.3390/futuretransp5020071>
- [8]. Çolak, S., Lima, A., & González, M. C. (2016). Understanding congested travel in urban areas. *Nature Communications*, 7, Article 10793. <https://doi.org/10.1038/ncomms10793>
- [9]. Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2), 224–227. <https://doi.org/10.1109/TPAMI.1979.4766909>
- [10]. Emenike, C. C., Bhero, E., & Twala, B. (2026). Optimised predictive modelling for highway traffic flow using real-world data. *Scientific African*, 32, Article e03330. <https://doi.org/10.1016/j.sciaf.2026.e03330>
- [11]. Faheem, H. B., El Shorbagy, A. M., & Gabr, M. E. (2024). Impact of traffic congestion on transportation system: Challenges and remediations – A review. *Mansoura Engineering Journal*, 49(2), Article 18.
- [12]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- [13]. Garber, N. J., & Hoel, L. A. (2019). *Traffic and highway engineering* (5th ed., Enhanced SI ed.). Cengage Learning.

- [14]. Greenshields, B. D. (1935). A study of traffic capacity. *Proceedings of the Highway Research Board*, 14, 448–477.
- [15]. Hammoumi, L., Farah, S., Benayad, M., Maanan, M., & Rhinane, H. (2025). Leveraging machine learning to predict traffic jams: Case study of Casablanca, Morocco. *Journal of Urban Management*, 14(3), 813–826.
- [16]. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- [17]. Helbing, D. (1997). Fundamentals of traffic flow. *Physical Review E*, 55(3), 3735–3742.
- [18]. Hirschman, A. O. (1964). The paternity of an index. *The American Economic Review*, 54(5), 761–762.
- [19]. Imoh, U. U., & Rad, M. M. (2025). Analysis and prediction of traffic conditions using machine learning models on Ikorodu Road in Lagos State, Nigeria. *Infrastructures*, 10(5), 122. <https://doi.org/10.3390/infrastructures10050122>
- [20]. Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137. <https://doi.org/10.1109/TIT.1982.1056489>
- [21]. Murisela, A. K., Olayinka, I. A., & John, W. (2025). Analysis of traffic-flow characteristics and vehicle composition using automatic traffic count data on the Lokoja–Benin road corridor. *EDSU Journal of Engineering and Technology*, 1(1), 1–15.
- [22]. Mystakidis, A., Koukaras, P., & Tjortjis, C. (2025). Advances in traffic congestion prediction: An overview of emerging techniques and methods. *Smart Cities*, 8(1), 25. <https://doi.org/10.3390/smartcities8010025>
- [23]. Nofal, S. (2024). Predicting vehicle travel time on city streets for trip preplanning and predicting heavy traffic for proactive control of street congestion. *Scientific Reports*, 14, Article 10497. <https://doi.org/10.1038/s41598-024-61379-7>
- [24]. 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](https://doi.org/10.1016/0377-0427(87)90125-7)
- [25]. Sokido, D. L. (2024). Measuring the level of urban traffic congestion for sustainable transportation in Addis Ababa, Ethiopia: The cases of selected intersections. *Frontiers in Sustainable Cities*, 6, Article 1366932.
- [26]. Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83. <https://doi.org/10.2307/3001968>
- [27]. Xing, Z., Huang, M., & Peng, D. (2023). Overview of machine-learning-based traffic-flow prediction. *Digital Transportation and Safety*, 2(1), 1–11. <https://doi.org/10.48130/DTS-2023-0013>
- [28]. Yao, J., He, R., Shi, T., Wang, P., & Zhao, X. (2023). Review on machine learning-based traffic flow prediction methods. *Journal of Traffic and Transportation Engineering*, 23(3), 44–67.
- [29]. Zhu, D., Ng, C., Xie, L., & Liu, Y. (2025). Interpretable machine learning for traffic congestion prediction: Unveiling the impact of different COVID-19 periods. *Communications in Transportation Research*, 5, Article 100226. <https://doi.org/10.1016/j.commtr.2025.100226>
- [30]. Zou, X., & Chung, E. (2024). Traffic prediction via clustering and deep transfer learning with limited data. *Computer-Aided Civil and Infrastructure Engineering*, 39(17), 2683–2700. <https://doi.org/10.1111/mice.13207>