Arterial Machine Learning Diagnostics (2025 CBI): Application of DBSACN and XGBoost to TMC-level congestion analysis

Abstract

The script ran four machine-learning checks on 8,681 arterial TMCs from the 2025 CBI. Together they produce short review lists, not corrections. Each check uses a different model, described below.

A. Data-quality outliers: robust z-scores and DBSCAN. Nine input features (volumes, the Replica hourly profile and count anchors) are each converted to a *robust z-score*. The score uses the median and the median absolute deviation, so a few extreme values cannot distort it. Each TMC is scored against its peers (same functional class and area type). *DBSCAN*, a density-based clustering method, then groups TMCs with similar z-score patterns. Its radius is set from a *nearest-neighbour* search (98th percentile of the distance to the 10th-nearest TMC, eps = 3.09). A TMC that falls in no dense group and has at least one $|z| > 3.5$ is flagged. Result: 67 TMCs flagged, and no count anchors outside 0.33–3.0.

B. Spatial hotspots: DBSCAN on map distance. TMCs congested on at least 50% of weekdays are placed at their midpoints. DBSCAN groups them using *haversine* (earth-surface) distance, with a radius of 0.5 km and at least 4 TMCs per core point. Result: 223 clusters covering 409 miles and holding 25% of all arterial delay. Most clusters (185) span more than one ranked section.

C. Drivers of congestion: gradient boosting with grouped cross-validation. A *gradient-boosting regressor* (hundreds of small decision trees, each correcting the last) predicts congestion intensity. Intensity is the volume-weighted excess travel-time ratio, which contains no AADT. The model uses 16 road characteristics and nothing derived from speeds. The design uses XGBoost, but this run used scikit-learn's HistGradientBoostingRegressor because xgboost was not installed. TMCs are weighted by length. Performance is measured with *grouped 5-fold cross-validation*: each road section stays in one fold. Feature influence is measured with *permutation importance*; SHAP values would replace it when XGBoost is used. Result: out-of-fold $R^2 = 0.535$ (MAE = 0.079). TMC length and traffic volume carry 39% of the importance.

D. Residual outliers: model residuals with robust z-scores. No new model is fitted. Each TMC's observed intensity is compared with its out-of-fold prediction from C. The gap (the residual) is robust-z-scored within its area-type band, and $|z| > 3$ is flagged. Result: 369 TMCs much more congested than similar roads and 73 much less. Short TMCs (0.1 mi or less) account for over half of these flags.

This write-up explains each method in plain language and walks through what this run found. Nothing in it changes the CBI results.

Contents

1 Summary at a glance **3**

2 The big picture: how the script flows **3**

3 Key ideas in plain words **4**

4 Step 0: loading the data and building features **5**

5 Analysis A: data-quality outliers **6**
5.1 How it works . . . 6
5.2 What it found . . . 7

6 Analysis B: spatial hotspot clusters **8**
6.1 How it works . . . 8
6.2 What it found . . . 9

7 Analysis C: what drives congestion intensity **11**
7.1 The question and the target . . . 11
7.2 Inputs . . . 12
7.3 How the model is trained and judged . . . 12
7.4 What it found . . . 13

8 Analysis D: roads that behave differently from their peers **14**
8.1 How it works . . . 14
8.2 More congested than expected . . . 15
8.3 Less congested than expected . . . 16

9 Things to keep in mind about this run **16**

10 Suggested next steps **17**

A Settings, output files and glossary **18**

List of Figures

1 Pipeline of the script . . . 4
2 DBSCAN on toy data . . . 5
3 Data-quality reasons per feature . . . 7
4 Hotspot cluster map . . . 10
5 Cluster size vs. delay . . . 11
6 Grouped cross-validation . . . 12
7 Feature importance . . . 13
8 Observed vs. predicted . . . 14
9 Residual z distribution . . . 15
10 Effect of TMC length . . . 17

1 Summary at a glance

Analysis	Question it answers	Method	Headline result
A. Data quality	Which TMCs have inputs (volumes, profiles) that look wrong?	Robust z-scores + DBSCAN	67 TMCs flagged; 0 count anchors outside 0.33–3.0
B. Spatial hotspots	Where do recurrently congested TMCs bunch together?	DBSCAN on map distance	223 clusters: 2,087 TMCs, 409 miles, 25% of all arterial EAVHD
C. Drivers	What road characteristics go with congestion intensity?	Gradient boosting + grouped cross-validation	Out-of-fold $R^2 = 0.535$; TMC length and volume lead
D. Residual outliers	Which TMCs are far more (or less) congested than similar roads?	Model residuals + robust z	369 much MORE congested, 73 much LESS

Table 1: The four analyses and their headline results.

Three things to know before reading further:

- This run used the scikit-learn fallback model because the `xgboost` package was not installed. There are no SHAP values: the importance table has no "direction", the SHAP dependence file is empty and the "top drivers" column is blank.
- Very short TMCs (0.1 mile or less) are 38% of all TMCs but supply 54% of the Analysis D flags. Their speeds are noisy.
- Every list is a place to start looking. A flag means "unusual", not "wrong".

2 The big picture: how the script flows

The script reads the finished CBI tables, builds one row of numbers ("features") per TMC, runs four analyses on that table, and writes CSVs, a summary and a map layer. It never writes into the CBI results; the optional `-write-db` flag only creates its own table, `"Year_2025".arterial_ml_diagnostics`.

Analyses A, B and C each read the same feature table independently. Analysis D is not a separate model: it reuses the predictions from C and looks at where they miss (Figure 1). In code the order is `load_data()` → `build_features()` → `analyse_quality()` (A) → `analyse_spatial()` (B) → `analyse_model()` (C and D) → writers.

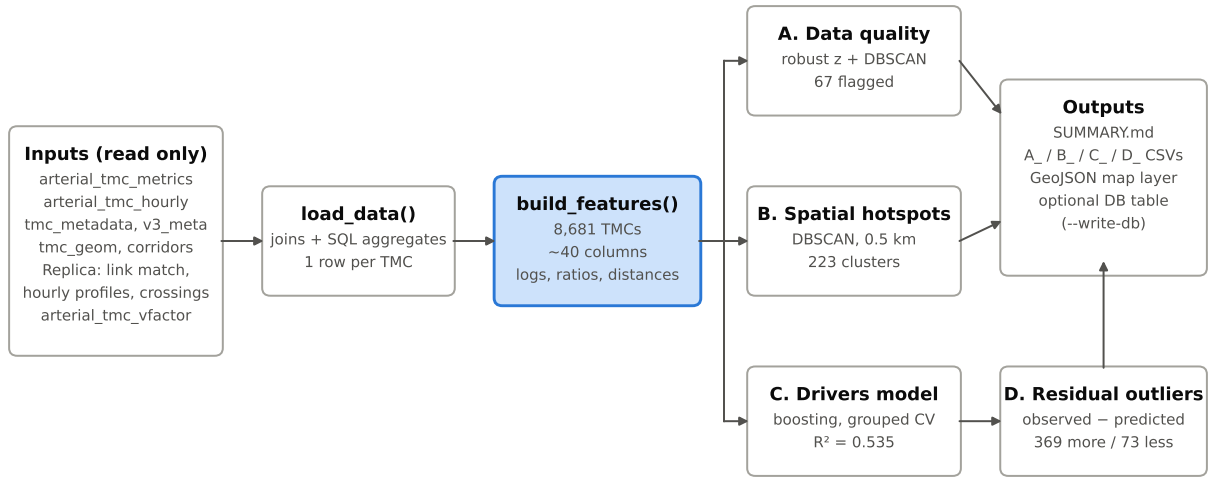

Figure 1: How the script flows, from read-only inputs to outputs.

3 Key ideas in plain words

TMC. A Traffic Message Channel segment: a short, one-direction piece of road with its own speed record in NPMRDS. Opposite directions of the same road are separate TMCs.

AADT and directional AADT. Average annual daily traffic (vehicles per day). Directional AADT is the part going in the TMC's direction. In the CBI it mostly comes from Replica, sometimes adjusted ("anchored") to a GDOT count.

Speed ratio and excess travel time. Speed ratio = observed speed ÷ free-flow speed. A ratio of 0.8 means the trip takes $1/0.8 = 1.25$ times as long: 25% excess travel time. The model target in Analysis C is a volume-weighted average of that excess over weekday hours 6 AM–9 PM:

$$y \;=\; \frac{\sum_h w_h \,\max\!\left(0,\; \dfrac{1}{\text{speed ratio}_h} - 1\right)}{\sum_h w_h} \tag{1}$$

where w_h is the Replica share of daily volume in hour h. A value of $y = 0.20$ means a daytime vehicle takes, on average, 20% longer than at free flow.

EAVHD. Excess annual vehicle-hours of delay, the CBI's headline delay measure: roughly excess travel time × volume, summed over the year. Big EAVHD needs both slow speeds and many vehicles.

Robust z-score. "How unusual is this value compared with its peers", in a way a few extreme values cannot distort. It uses the median and the median absolute deviation (MAD) instead of the mean and standard deviation:

$$z \;=\; \frac{0.6745\,(x - \text{median})}{\text{MAD}}, \qquad \text{MAD} = \text{median}\big(|x - \text{median}\,|\big) \tag{2}$$

The factor 0.6745 makes it comparable to an ordinary z-score for normal data, so $z = 3.5$ is about "3.5 standard deviations from typical". Peer groups with fewer than 20 values borrow the overall median and MAD.

DBSCAN. A clustering method that finds dense groups without being told how many exist. It has two settings: a radius ε (eps) and a minimum count (min_samples). A point with at least min_samples points within eps (itself included) is a *core* point; points near a core point join its cluster (*border* points); points near no core point are *noise* (Figure 2). The script uses DBSCAN twice: in A, "noise" means a TMC whose input profile resembles no dense group; in B, clusters are places on the map.

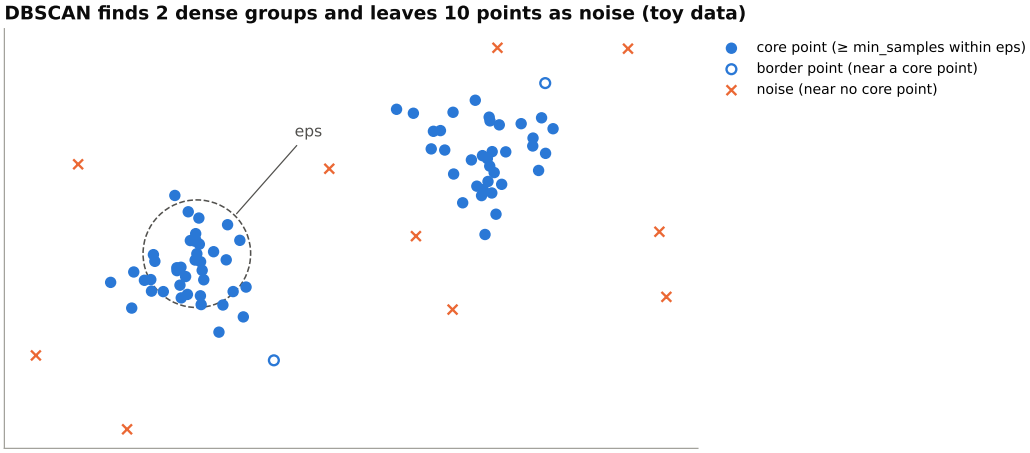

Figure 2: DBSCAN on made-up points: two dense groups are found and scattered points are left as noise.

Gradient boosting. A prediction model built from hundreds of small decision trees; each new tree corrects the errors of the trees before it. It captures non-linear effects and interactions ("high volume matters more in the CBD") without you writing them down.

Cross-validation and out-of-fold predictions. To judge a model fairly, test it on data it did not learn from. The data are split into 5 folds; the model trains on 4 and predicts the 5th, five times over. Each TMC then has one out-of-fold prediction from a model that never saw it. R^2 on these predictions is an honest score (1 = perfect, 0 = no better than predicting the average).

Feature importance. How much a model relies on each input. SHAP (with xgboost) splits every prediction into per-feature contributions. Permutation importance (used in this run) shuffles one feature at a time and measures how much worse the predictions get.

4 Step 0: loading the data and building features

The script turns several database tables into one table with 8,681 rows (one per analysed arterial TMC) and about 40 columns. All later steps read only this table.

Piece	Source table(s)	What it adds
Base row	`arterial_tmc_metrics`, `tmc_metadata`, `v3_meta`	road, section, county, area type, length, AADT, directional AADT and source text, EAVHD, occurrence %, functional class, NPMRDS lanes and AADT
Profile shape	`arterial_tmc_vfactor`	weekend/weekday ratio, AM share (7–10 AM), PM share (4–7 PM), night share (9 PM–6 AM), peak-hour share
Target	`arterial_tmc_hourly` + weekday factors	excess travel-time ratio y and daytime hours with speeds
Lanes (backup)	`tmc_link_match`, `hourly_link_profile`	OSM/Replica directional lanes, weighted by match share
Crossings	`tmc_link_candidates` (CROSSING)	counts of major and of all crossing streets
Freeway distance	`tmc_geom` + active Interstate/Expressway corridors	metres to nearest freeway TMC (PostGIS)

Table 2: Where each piece of the feature table comes from (`load_data`).

In this run the profile source covered all arterial TMCs and all 8,681 TMCs had at least 10 daytime hours of speeds, so none were dropped from the model. The derived features (`build_features`) are:

- **Logs.** Volumes and ratios are logged. On a log scale "half" and "double" are the same distance from 1.
- **Count-anchor factor.** Parsed from the text in `dir_aadt_source`: a source ending in "x1.92)" gives 1.92. 769 TMCs have one.
- **Neighbour jump.** Along each road and direction, the TMC's AADT ÷ the median of the TMCs before and after it.
- **Directional lanes.** NPMRDS through lanes (halved on two-way facilities), else OSM lanes; at least 1. AADT per lane follows.
- **Crossings per mile.** Crossing-street counts ÷ TMC length: an intersection-density proxy.
- **Distances.** To the nearest freeway and to Five Points (straight line from the TMC midpoint), in km.
- **Codes.** County as a number; "two-way profile" = 1 when the volume method is `REPLICA_COMB_REVIEWED`.

Nothing derived from speeds is a model input, so the model cannot "cheat" by seeing the answer.

5 Analysis A: data-quality outliers

Analysis A flagged **67 TMCs** whose inputs look unlike any normal group, and found **no** count anchor outside 0.33–3.0. It looks only at inputs, never at congestion, so a flag here is about data, not traffic.

5.1 How it works

1. **Nine input features** that would look strange if a volume or profile were wrong: directional AADT, Replica vs. NPMRDS AADT, count-anchor factor, neighbour jump, AADT per lane,

weekend/weekday ratio, peak-hour share, night share, and share of volume in hours with no speed data.

2. **Score against peers.** A robust z-score per feature inside the peer group (functional class × area band).
3. **DBSCAN on the nine z-scores.** Missing scores become 0 and scores are capped at ±10. eps is the 98th percentile of each TMC's distance to its 10th-nearest neighbour; here eps = 3.09, min_samples = 10.
4. **Keep noise TMCs with a clear reason:** DBSCAN noise *and* at least one $|z| > 3.5$. Those features are the reasons.
5. **Count anchors** below 0.33 or above 3.0 are listed separately.

Why two filters? 1,148 TMCs have at least one feature beyond 3.5. DBSCAN narrows that to TMCs unusual *overall*. In this run all 67 noise TMCs also had a reason.

5.2 What it found

The neighbour jump is the most common oddity, among all TMCs and among the flagged (Figure 3).

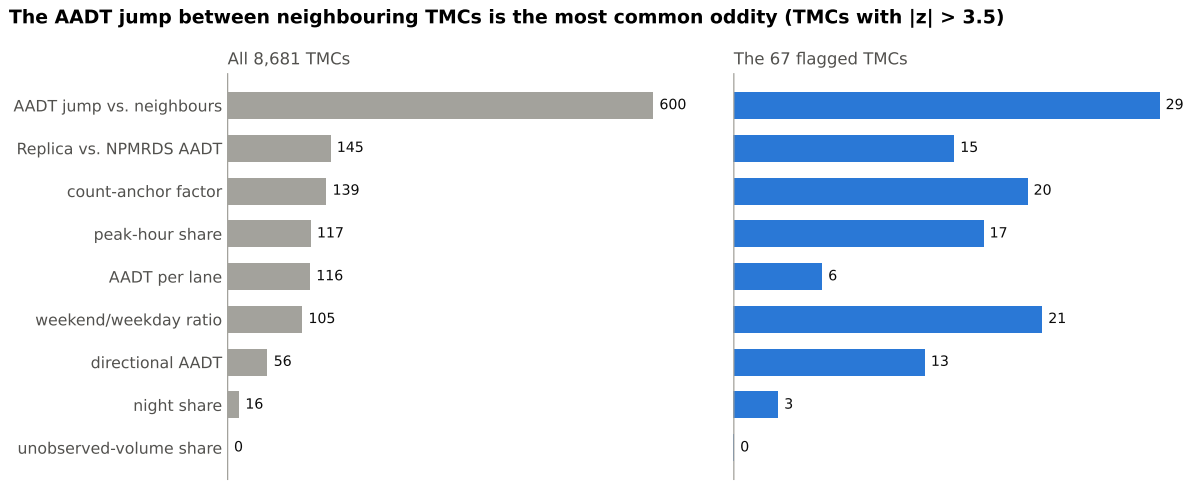

Figure 3: TMCs with $|z| > 3.5$ per feature: all TMCs (left) and the 67 flagged (right).

Pattern	Example	What it suggests
Very low volume next to much higher neighbours	GA-139 / W Point Ave at Camp Creek Pkwy: 335 veh/day, jump $z = -22$	matched to a minor Replica link ("OSM tertiary, full file")
Odd hourly profile	Klondike Rd SW: peak-hour share $z = +18.5$; Windy Hill Rd: weekend ratio $z = -6.8$, peak share $z = +13.6$	a school, plant or event generator, or a bad profile match
Count anchor far from typical	US-278 at E Lake Dr: anchor $z = -15$, AADT ≈1,170	check the count station and its direction
Replica and NPMRDS disagree	GA-9 at GA-53: $z = -12.5$; GA-166 at GA-5: $z = +11.6$	one of the two AADT sources is off

Table 3: Four patterns among the 67 flagged TMCs.

Count-anchored TMCs are 9% of all TMCs but 39% of the flagged (26 of 67); "OSM tertiary" matches are 2% of all but 13% of the flagged (9 of 67). The flagged TMCs carry only 0.3% of total EAVHD. All 769 anchor factors lie between 0.33 and 2.77; the lowest sits exactly on the 0.33 limit and is not listed because the test is strict. The "unobserved volume share" feature is zero for nearly every TMC, so its MAD is zero and it had no effect.

6 Analysis B: spatial hotspot clusters

Analysis B found **223 clusters**: 409 miles of arterial holding 25% of all arterial EAVHD (32.9 of 131.2 million vehicle-hours). Clusters can cross roads: 185 of 223 include more than one ranked section.

6.1 How it works

1. Keep TMCs congested on at least 50% of weekdays: 3,504 TMCs.
2. Place each at its midpoint.
3. DBSCAN with haversine (earth-surface) distance, eps = 0.5 km, min_samples = 4.
4. Summarise each cluster (TMCs, miles, EAVHD and per mile, sections and their shares, counties, area band, share near a freeway, mean occurrence).
5. Rank by total EAVHD.

Of the 3,504 congested TMCs, 2,087 joined a cluster; 1,417 are isolated.

6.2 What it found

Rank	Area (top sections)	TMCs	Miles	EAVHD	per mile
1	W Midtown/Downtown: Northside Dr, Ivan Allen Jr Blvd, Marietta St	168	23.9	1,733,203	72,403
2	Buckhead: Peachtree Rd, Piedmont Rd, Lenox Rd	48	7.9	1,165,249	147,224
3	Lindbergh: Piedmont Rd, Cheshire Bridge Rd, Sidney Marcus Blvd	40	7.2	922,457	128,458
4	Midtown: Peachtree St, 10th St, Spring St, 14th St	62	7.2	921,715	128,051
5	Duluth: Pleasant Hill Rd, Steve Reynolds Blvd, Satellite Blvd	25	6.3	842,071	133,330
6	Midtown south: North Ave, Spring St, Juniper St, Monroe Dr	43	7.3	793,323	108,911
7	Norcross: Jimmy Carter Blvd, Norcross Tucker Rd	18	4.7	730,459	155,624
8	Clayton: Tara Blvd (GA-54 / US-19)	27	6.6	702,554	105,755
9	McDonough Rd (GA-20 / GA-138)	4	4.4	687,895	156,525
10	Gainesville: Jesse Jewell Pkwy, Riverside Dr, E. E. Butler Pkwy	28	6.6	559,734	85,014

Table 4: The ten largest hotspot clusters by annual delay (EAVHD in vehicle-hours).

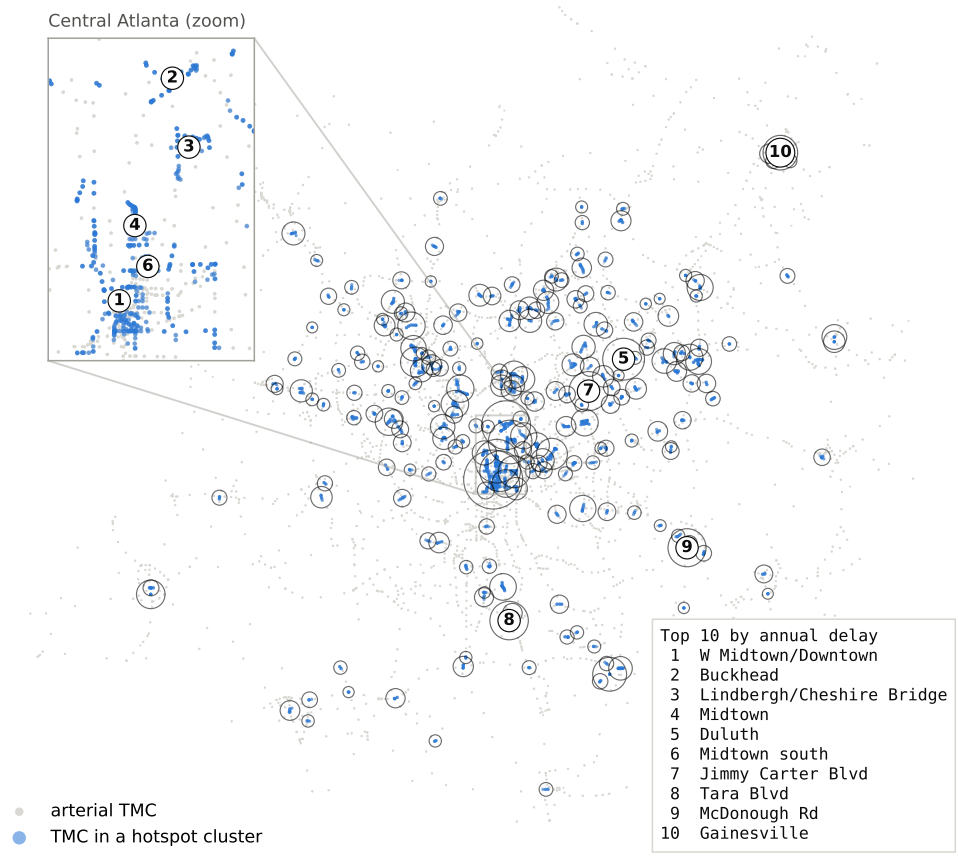

Figure 4: TMC midpoints (grey), TMCs in a cluster (blue) and cluster delay (circles).

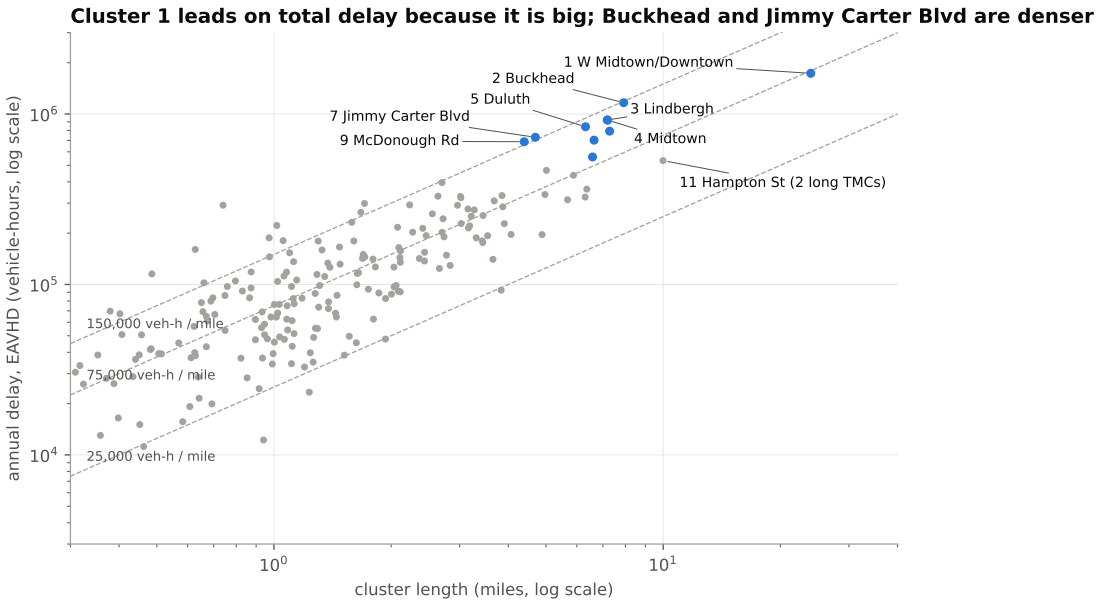

Figure 5: Each dot is a cluster; dashed lines mark equal delay per mile. Blue = top 10.

- **Total vs. intensity.** Cluster 1 leads because it is big; per mile the densest are McDonough Rd (156,525), Jimmy Carter Blvd (155,624) and Buckhead (147,224).
- **Where.** 129 suburban, 70 urban, 20 CBD, 4 exurban/rural. By county: Fulton 46, Cobb 45, Gwinnett 39, DeKalb 29.
- **Concentration.** The top 10 clusters hold 6.9% of all arterial delay.
- **Both directions count.** Opposite directions share a midpoint, so a 4-TMC cluster can be just two places.
- **Long TMCs stretch the rule.** Rank 11 (Hampton St) is 4 TMCs but 10 miles: two long TMCs meeting near a point.
- The area band is the cluster's most common one; rank 5 (Duluth) shows "CBD" because of area type 1 in the source data.

7 Analysis C: what drives congestion intensity

Using road characteristics alone, the model explains about half the variation in congestion intensity: out-of-fold $R^2 = 0.535$, MAE = 0.079. TMC length and traffic volume carry about 39% of total importance.

7.1 The question and the target

"Given what a road looks like, how congested should it be?" The answer is y from Equation 1. Its length-weighted mean is 0.196; half of TMCs are below 0.21 and 1 in 10 is above 0.59. The target is deliberately not EAVHD: EAVHD already contains AADT, so predicting it from AADT would just relearn the CBI formula.

7.2 Inputs

Sixteen features, none from speeds: log directional AADT, AADT per lane, directional lanes, functional class, area type, TMC length, major and all crossings per mile, distance to freeway, distance to downtown, weekend/weekday ratio, AM share, PM share, peak-hour share, two-way profile flag, county.

7.3 How the model is trained and judged

1. **Model.** With xgboost: 600 trees, depth 5, learning rate 0.03, 80% row/column sampling. Without it (this run): `HistGradientBoostingRegressor`, 600 iterations, learning rate 0.03, $\leq$31 leaves, $\geq$20 TMCs per leaf, L2 = 1.0, county as a category.
2. **Weights.** Each TMC counts in proportion to its length (minimum 0.01 mi).
3. **Grouped 5-fold cross-validation** by section (`arterial_group`), so neighbouring TMCs of one road never sit on both sides of a split (Figure 6).
4. **Score.** Length-weighted R^2 and MAE on out-of-fold predictions.
5. **Importance.** A final model on all TMCs; permutation importance (5 repeats) in this run, mean |SHAP| with xgboost.

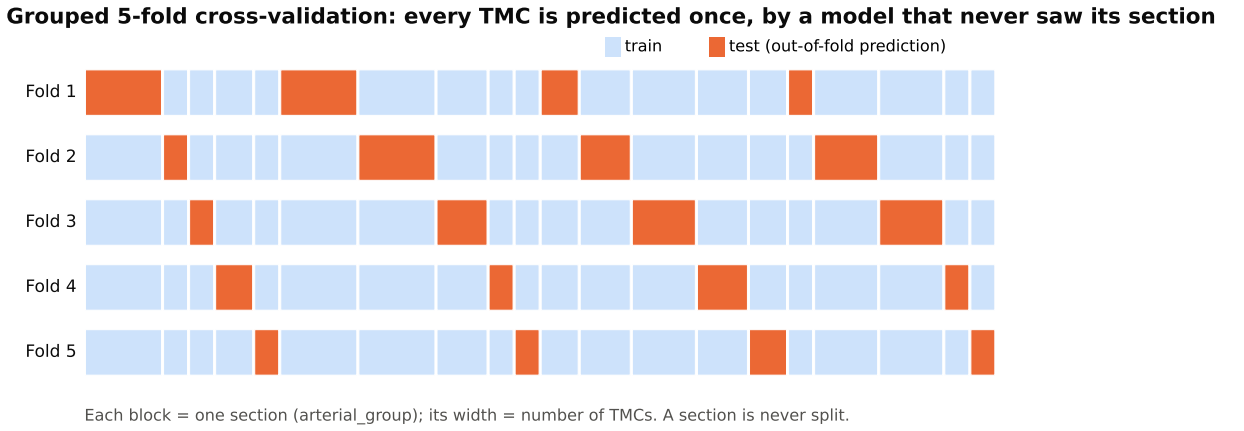

Figure 6: Grouped 5-fold cross-validation, illustrated with 20 made-up sections.

7.4 What it found

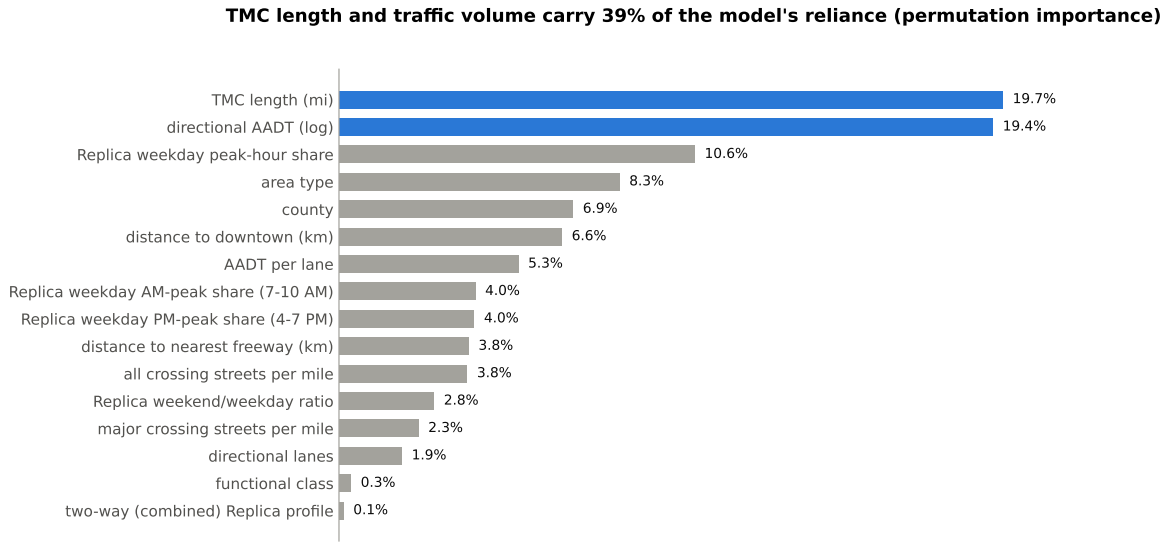

Figure 7: Permutation importance as a share of the total.

- **TMC length (19.7%).** Short TMCs are more congested per vehicle; a TMC of a few hundred feet often sits right at a signal. Partly real (approaches are where delay happens), partly how TMCs are cut.
- **Directional AADT (19.4%).** More traffic, more delay per vehicle.
- **Peak-hour share (10.6%).** Traffic piled into one hour gives sharper peaks.
- **Area type (8.3%), county (6.9%), distance to downtown (6.6%).** Location: density, signals, turning, pedestrians.
- **AADT per lane (5.3%).** Lower than plain AADT, possibly because lane counts are noisy.
- **Crossings per mile (3.8%, 2.3%)**; **functional class (0.3%)** and the **two-way flag (0.1%)** add little.

Permutation importance has no direction; the directions above are the usual traffic-engineering reading, to be confirmed with SHAP. For a model that never sees a speed, $R^2 = 0.535$ is solid; the rest is local (signals, access, geometry, incidents) plus noise. Unweighted, $R^2 = 0.364$: the model is weaker on short TMCs.

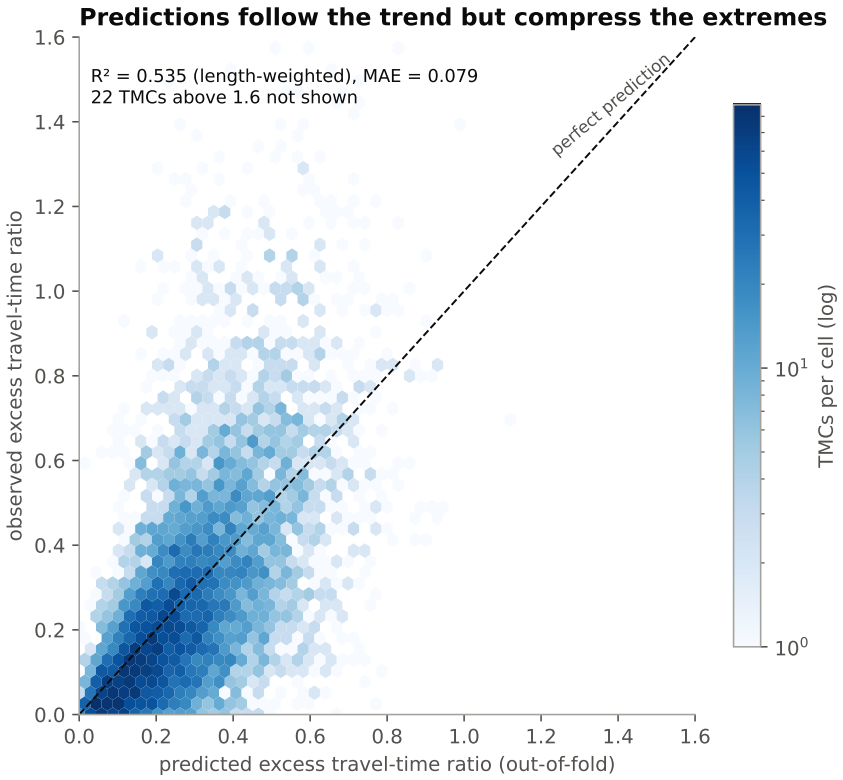

Figure 8: Observed vs. out-of-fold predicted y. Points above the line are more congested than predicted.

8 Analysis D: roads that behave differently from their peers

Analysis D flagged **369** TMCs as much MORE congested than similar roads and **73** as much LESS. Together they carry 8.1% of all arterial EAVHD.

8.1 How it works

1. **Residual** = observed y − out-of-fold predicted y; the model never saw that TMC's section.
2. **Robust z of the residual** within each area band, so CBD streets do not crowd out the rest.
3. **Flag** $z > 3$ (MORE congested) and $z < -3$ (LESS congested).
4. **Top drivers** (xgboost only) — blank in this run.

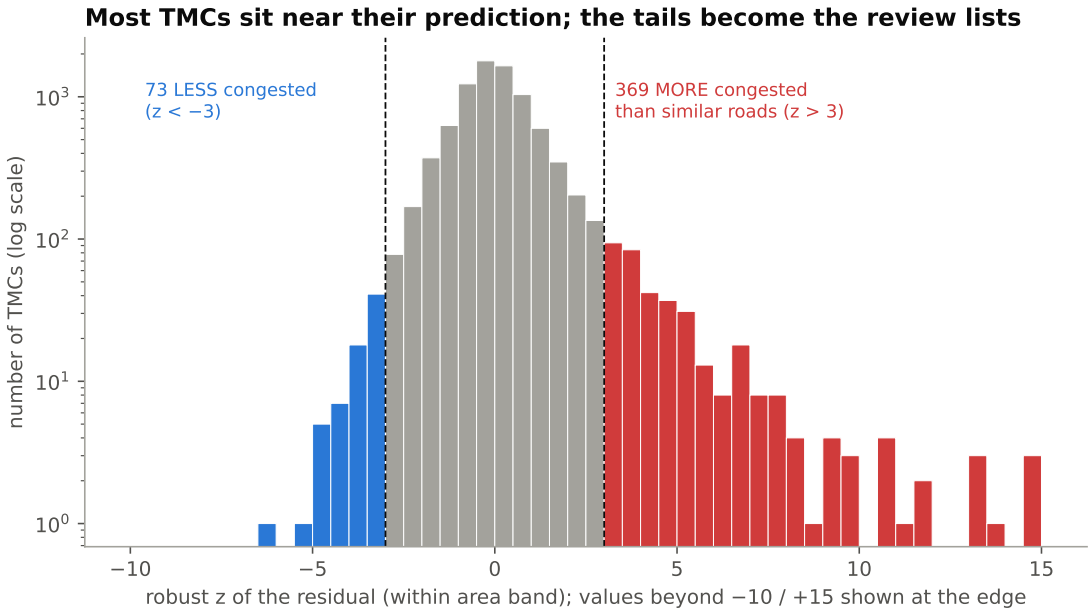

Figure 9: Residual z-scores; red and blue bars are the flagged tails.

Suburban TMCs account for 224 of the 369 MORE flags; exurban/rural has the highest rate of LESS flags (22 of 932).

8.2 More congested than expected

TMC	Road at	Obs. y	Pred. y	z	EAVHD
101-05265	GA-140 at US-19/GA-400	1.08	0.50	3.7	361,068
101-07424	GA-124 at Ronald Reagan Pkwy	0.84	0.22	5.3	265,709
101+10965	GA-54 at GA-74	0.52	0.26	4.6	223,576
101+07843	GA-124 at Webb Gin House Rd	0.76	0.29	4.0	211,323
101+11170	GA-138 at US-278	0.88	0.53	3.1	199,808
101-05158	Pleasant Hill Rd at Club Dr	1.07	0.71	3.1	196,721
101-05148	GA-316 at US-29 (Winder Hwy)	0.88	0.31	9.8	195,291
101-05579	US-23 at Memorial Dr	1.03	0.41	4.0	186,626

Table 5: Top MORE-congested TMCs by annual delay.

On GA-140 approaching GA-400, daytime trips take 2.08 times free-flow time ($y = 1.08$); similar roads typically see $y = 0.50$. GA-124 appears three times in the top ten; GA-120 and GA-20 each appear several times in the top 25. The largest z-scores belong to very short TMCs: GA-166 (Maple St, Carroll County) is 0.05 mi long with $y = 2.40$ against a prediction of 0.14 ($z = 38$).

8.3 Less congested than expected

TMC	Road at	Obs. y	Pred. y	z	EAVHD
101+11717	GA-120-ALT at Church St	0.71	1.13	−3.2	54,242
101-07084	Sugarloaf Pkwy at Old Snellville Hwy	0.46	0.92	−3.5	46,436
101+07089	Sugarloaf Pkwy at Cruse Rd	0.26	0.76	−4.0	20,506
101-12853	Courtland St at Ellis St	0.09	0.79	−3.8	14,777
101N10588	17th St NW at I-85/I-75	0.13	0.71	−3.2	9,217

Table 6: Top LESS-congested TMCs by annual delay.

Sugarloaf Pkwy (three entries) is built like a parkway with few driveways; downtown one-way streets often have coordinated signals; 17th St crosses the Downtown Connector on a bridge. The model knows nothing of access control or signal coordination, so it expects them to be worse. Only 3 of the 442 flagged TMCs are also data-quality outliers.

9 Things to keep in mind about this run

1. **No SHAP.** The fallback model ran; predictions are fine, explanations are thinner.
2. **Short TMCs are noisy** (Figure 10). 3,264 TMCs (38%) are 0.1 mi or shorter; their average miss is three times that of TMCs over 2 mi (0.16 vs. 0.05) and they hold 240 of the 442 D flags.
3. **Length weighting.** $R^2 = 0.535$ counts road-miles; per TMC it is 0.364.
4. **Importance on training data.** Features the model can memorise (county, length) may look somewhat more important; related features (AADT and AADT per lane; AM, PM and peak shares) share credit.
5. **Counts depend on settings.** eps in A is a percentile, so some noise always appears; likewise 0.5 km and 50% in B and 3.0 in D.
6. **Two small input issues:** the zero-MAD feature, and the strict anchor test at 0.33.

Above all: associations, not causes.

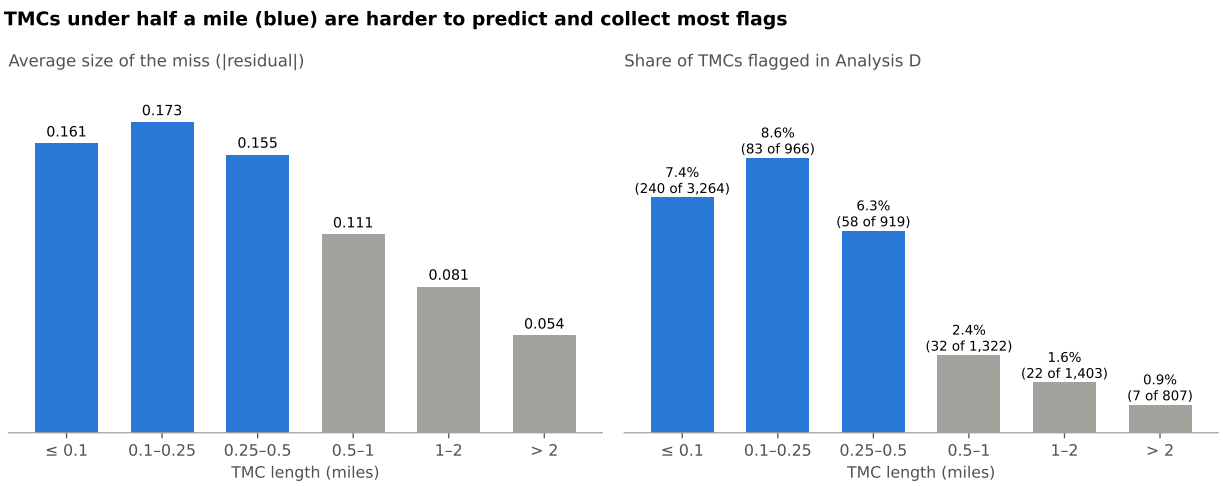

Figure 10: Average miss and share flagged in Analysis D, by TMC length.

10 Suggested next steps

1. Install xgboost (`pip install xgboost`) and rerun for directions, dependence and per-TMC drivers.
2. Review the 67 quality flags, starting with the 26 count-anchored TMCs and the 9 "OSM tertiary" matches.
3. Filter D to TMCs ≥ 0.25 mi or roll up to sections; start with GA-124, GA-120, GA-20 and GA-140 at GA-400.
4. Use B clusters for area-wide studies (Buckhead, Midtown, Duluth, Jimmy Carter Blvd).
5. Small code fixes: an "any unobserved volume" indicator, an inclusive anchor test (≤ 0.33), optionally one point per road location in B.

A Settings, output files and glossary

Setting	Value	Used in	Meaning
Z_FLAG	3.5	A	$\|z\|$ to name a feature as a reason
QC_MIN_SAMPLES	10	A	DBSCAN min_samples on z-scores
QC_EPS_QUANTILE	0.98	A	eps = 98th pct of 10th-neighbour distance (3.09)
ANCHOR_MIN/MAX	0.33, 3.0	A	plausible anchor range
CONGESTED_OCC	50%	B	weekdays congested to enter clustering
SPATIAL_EPS_KM	0.5	B	DBSCAN radius (km)
SPATIAL_MIN_SAMPLES	4	B	TMCs around a core point
NEAR_FREEWAY_M	500	B	"near a freeway" (m)
DAY_HOURS	6–20	C	weekday hours 6 AM–9 PM
MIN_TARGET_HOURS	10	C	daytime hours with speeds needed
N_FOLDS	5	C	cross-validation folds
RESID_Z	3.0	D	residual z cut-off
RANDOM_STATE	42	C	repeatability

Table 7: Settings used in this run.

File	Rows	Contents
SUMMARY.md	–	short report the script prints
A_quality_all_tmcs.csv	8,681	every TMC with nine z-scores, score, reasons
A_quality_outliers.csv	67	flagged TMCs
A_count_anchor_extremes.csv	0	anchors outside 0.33–3.0
B_spatial_clusters.csv	223	one row per cluster
B_spatial_cluster_members.csv	8,681	cluster rank per TMC (-1 = none)
C_feature_importance.csv	16	importance per feature
C_shap_dependence.csv	0	needs xgboost
D_model_predictions_all.csv	8,681	observed, predicted, residual, z
D_residual_outliers.csv	442	flagged TMCs by z
arterial_ml_diagnostics.geojson	8,681	map layer with flags and cluster ranks

Table 8: Output files.

Term	Meaning
Core / border / noise	DBSCAN roles: dense centre, edge of a cluster, belongs to none
eps	DBSCAN radius
Haversine	distance on the earth's surface between two latitude/longitude points
MAD	median absolute deviation
MAE	mean absolute error: average size of a miss
Out-of-fold	predicted by a model that did not train on that row
Peer group	TMCs compared with each other (class × area band in A; area band in D)
Permutation importance	loss in accuracy when one feature is shuffled
R^2	share of variation explained (1 perfect, 0 = average)
Residual	observed minus predicted
SHAP	split of one prediction into per-feature contributions

Table 9: Glossary.