DOI: **10.37943/25VWNB3103**

Bakbergen Mendaliyev
PhD student, Department of Computer Engineering
cornix1999@gmail.com, orcid.org/0009-0003-5529-3322
Astana IT University, Kazakhstan
Didar Yedilkhan
Doctor of Philosophy, Department of Computer Engineering
d.yedilkhan@astanait.edu.kz, orcid.org/0000-0002-6343-5277
Astana IT University, Kazakhstan
Aidarbek Shalakhmetov
PhD student, Department of Computer Engineering
aidar.shalakhmetov@gmail.com, orcid.org/0009-0001-6779-770X
Astana IT University, Kazakhstan

DETECTING AND RANKING URBAN BOTTLENECKS FROM PASSIVE SPEED AGGREGATES

Abstract: Urban traffic congestion remains a persistent problem, yet many cities still lack dense sensor networks, calibrated simulation models, or detailed origin-destination data for operational bottleneck monitoring. This study develops a lightweight framework for detecting and ranking urban bottlenecks using passive probe-based speed aggregates alone. For each road segment, a free-flow benchmark is estimated from high night-time speeds. Hourly median speeds are then converted into travel times, and cumulative delay is normalized by segment length and mildly regularized to reduce instability on short urban links. The framework is applied to a 20-day probe-data sample for Astana containing 6.34 million link-hour observations across 22,333 segments. After quality checks and coverage filtering, 8,634 segments remain in the final analysis set.

The main results are as follows:

- The estimated free-flow benchmark aligns closely with posted speed limits and remains stable under alternative percentile and night-window definitions.
- Congestion is strongly concentrated: 1,336 segments account for about 50% of the total delay, while 3,868 segments account for 80%, indicating a pronounced Pareto-type structure.
- The ranking remains robust after excluding days affected by major external disturbances, which suggests that the main bottleneck pattern is not driven by a small number of atypical days.
- Spatial diagnostics reveal significant positive autocorrelation in congestion severity, and the clustering pattern becomes stronger when road connectivity is represented with a network-based weight matrix rather than a purely geometric nearest-neighbour specification.
- Local cluster analysis identifies corridor cores of severe delay together with adjacent transition links, showing that the most critical bottlenecks are spatially connected rather than randomly scattered.
- Clustering of normalized 24-hour delay profiles reveals three evening-oriented regimes that differ mainly in congestion intensity.

Taken together, these findings show that routinely collected passive probe data can recover meaningful and operationally useful congestion structure even when a city lacks dense fixed-sensor coverage or a calibrated simulation model. The proposed workflow is transparent, reproducible, and suitable for corridor prioritization, before-and-after evaluation, and future digital-twin-based traffic management.

Keywords: traffic bottlenecks; passive speed aggregates; delay per kilometre; spatial autocorrelation; diurnal clustering; digital twin; urban congestion diagnostics.

Received: 28.11.2025 Accepted: 25.02.2026 Published: 30.03.2026

Introduction

Urban traffic congestion remains a persistent problem in many cities, especially where road networks already operate close to their practical capacity. Signal-related queues, minor incidents, and day-to-day demand fluctuations gradually reduce travel-time reliability for both commuters and freight. At the same time, city authorities still need to decide where interventions should be prioritized first. In practice, however, such decisions are often made without dense sensor coverage, current origin-destination matrices, or the time and resources required for detailed simulation studies. As a result, prioritization is frequently based on simple averages, public complaints, or isolated special studies. These inputs can misrepresent the actual bottlenecks and obscure how persistent some problem locations remain over time.

One increasingly accessible alternative is the use of passive speed aggregates: hourly segment-level statistics derived from anonymous GPS traces collected through navigation applications, connected vehicles, and smartphones. Such data are already available at city scale from several providers and can be updated regularly. Yet translating these aggregates into a defensible bottleneck ranking is not straightforward. Link lengths differ substantially, so naïve delay measures often overstate the importance of very short segments. Hourly medians also fluctuate from day to day, which means that a ranking based on a short observation period may not be stable. In addition, congestion is inherently spatial: interventions are usually aimed at corridors or neighbourhoods rather than isolated points. A credible workflow must therefore account for link length, assess temporal stability, and test whether the most affected links form coherent spatial patterns. Together, these challenges motivate the stepwise workflow proposed in this study.

In this study, we develop a minimal-data method that turns passive speed aggregates into practical evidence for bottleneck identification. The procedure is intentionally designed to remain simple enough for replication by non-specialist teams. Its main stages are as follows:

1. A conservative free-flow benchmark is estimated for each segment from night-time speeds.
2. Hourly median speeds are converted into travel times, and delay relative to the free-flow baseline is calculated.
3. Severity is expressed as delay per kilometre so that links of different lengths can be compared more fairly.
4. A mild length-based regularization is introduced to reduce volatility among short urban segments.
5. Ranking stability is examined under alternative parameter settings over a continuous twenty-day observation window, used here as a practical test of temporal consistency.
6. Spatial structure is analysed using spatial-statistical diagnostics and maps, while daily delay profiles are clustered to connect spatial and temporal bottleneck patterns.

The method is intended for settings where data are limited: observation windows are short, sampling is uneven, and link geometry is highly heterogeneous. Even under these constraints, the workflow produces tables and maps that can be independently checked and discussed in planning practice. The central question of the study is therefore straightforward: can passive speed aggregates alone support a city-wide bottleneck ranking that is fair across link lengths, stable over time, and spatially coherent enough to guide corridor-level interventions? We address this question using twenty consecutive days of segment-hour observations from Astana and present both the resulting rankings and the supporting diagnostics, with the aim of making the outputs directly usable for prioritization, before-and-after evaluation, and future digital-twin-based traffic management.

Literature Review

Recent work on congestion analysis converges on a common premise: bottlenecks should be identified from temporal and spatial patterns that remain robust under noise, rather than from simple averages alone. At the network scale, Serok, Havlin, and Lieberthal show that ranking

congestion “sources” by their system-wide cost can generate practical and ordered intervention lists for agencies [1]. At a much finer spatial resolution, trajectory-based studies isolate maneuver-specific hot spots at individual turns and forks, demonstrating that detailed vehicle movements help explain highly localized causes of delay [2]. Our study is positioned between these two ends of the spectrum. Rather than focusing either on network-wide system costs or on movement-level microdiagnostics, we ask whether passive segment-level speed aggregates alone can support a city-scale bottleneck ranking that remains comparable across locations and reasonably stable over time.

Several studies based on floating cars and crowdsourced speed data show that passive measurements can capture how congestion emerges, persists, and propagates across metropolitan networks [3]. They are able to reveal spillbacks and recurrent hot spots over extended periods. When flow estimates are needed, passive speeds have also been combined with learning-based methods to reconstruct large-scale freeway flows [4]. That direction is promising, but it also requires more calibration, more modelling effort, and more auxiliary data. Many agencies cannot meet these requirements quickly, especially when sensor infrastructure is sparse or analytic capacity is limited. In such settings, a more realistic objective is not to rebuild full origin-destination dynamics, but to derive a defensible congestion severity measure directly from speed aggregates and then test its usefulness through stability checks and spatial diagnostics.

Reliability has become an increasingly important lens for interpreting congestion. Decomposing travel times into a long-run trend, a regular daily cycle, and a residual component helps distinguish structural pressure from short-term noise and external shocks [5]. City-level performance indicators also show repeatable temporal patterns, including weekday-weekend contrasts, asymmetry between morning and evening peaks, and spikes linked to specific events [6]. These findings suggest that a credible ranking of critical locations should not change drastically when the observation window shifts by a few days. This naturally motivates rank-based evaluation. The most problematic locations should remain near the top under reasonable perturbations, and the degree of rank stability should be reported explicitly rather than assumed.

Data quality and spatial coverage remain recurring concerns in the passive-data literature. Cross-provider comparisons of travel times against Bluetooth-based references show broadly consistent trends, but also important differences in noise levels and geographic completeness [7]. These findings support the use of explicit thresholds on observation counts and consistent temporal aggregation schemes in any workflow based on passive aggregates. Case studies of community-reported or crowdsourced speeds show that these data become much more useful after calibration and cleaning [8]. Without such steps, short-lived spikes can easily appear as artificial congestion hot spots. Forecasting studies reach a similar conclusion: gains in performance often depend as much on disciplined feature engineering and data hygiene as on the predictive model itself [9]. Other corridor-level studies show that weather and incidents can shift travel times in systematic ways [10–11]. In the context of the present study, these factors are not modeled separately. Instead, they motivate robustness checks designed to assess whether bottleneck rankings remain broadly stable when major external disturbances are excluded.

Passive traffic analytics also connect congestion analysis with safety and operational planning. Probe-based speed indicators have been used to attribute part of crash risk to speeding behaviour across extensive road networks [12], which supports speed management as a possible co-benefit of congestion-mitigation strategies. Probe counts can also substitute for fixed detectors in microscopic simulation and provide high temporal resolution for operational assessment [13]. This is particularly relevant for before-and-after analysis, since it allows agencies to evaluate interventions without installing new hardware. In addition, studies of crowdsourced user feedback suggest that perceived trouble spots often coincide with crash hot spots [14]. This implies that subjective complaints and infrastructure bottlenecks often overlap and may be more informative when interpreted together rather than separately.

On the modelling side, several strands of work emphasize interpretability and parsimonious representation. Discrete-choice formulations that describe regime switching between free-flow, unstable, and congested states provide a narrative that practitioners can follow [15]. They make it easier to explain where pressure builds up, when it spills over, and how it dissipates. Similarly, simple and transparent travel-time prediction baselines provide useful reference points against which more complex models must justify their added value [16]. Automated feature-selection studies also show that dimensionality can often be reduced substantially without harming predictive performance [17]. This lesson is directly relevant to bottleneck ranking in data-sparse settings: when inputs are noisy and incomplete, compact and interpretable metrics may be more useful in practice than highly parameterized models.

Position papers on urban traffic management increasingly argue for a bottleneck-focused perspective [18]. They emphasize early identification of critical locations, explicit valuation of delay, and spatially aware prioritization at corridor scale. More recent studies extend this argument. Jin et al. propose a propagation-based method for identifying critical roads whose congestion effects spread to neighbouring links [19], showing that bottlenecks can be understood as network-level phenomena rather than isolated local failures. Zhang et al. introduce a resilience-oriented framework for ranking critical urban roads under sustainable mobility constraints [20], highlighting the interaction between network vulnerability and long-term planning priorities. Together, these contributions reinforce the idea that bottlenecks should be assessed not only by their own delay, but also by their position within broader spatial structure.

Despite this growing literature, an important gap remains. There is still limited evidence on whether a city-wide bottleneck ranking can be derived directly from passive segment-level speed aggregates in a way that is simultaneously length-aware, temporally stable, spatially coherent, and practical under short observation windows. Much of the existing work either relies on richer sensing environments, focuses on forecasting rather than diagnosis, or targets either highly localized maneuvers or broader network propagation effects. By contrast, our study develops a ranking-first diagnostic workflow for data-constrained settings. It does not require dense fixed sensors, calibrated microsimulation, or full origin-destination reconstruction. Instead, it combines passive speed aggregates with a length-aware severity metric, explicit stability checks, and spatial diagnostics to produce ranked shortlists and maps that can be used directly for intervention prioritization and before-after evaluation.

Three working principles follow from this literature and guide our design:

1. Severity should be normalized by exposure, including segment length, so that links of different sizes and functional classes can be compared more fairly.
2. Stability should be treated as an empirical outcome, which means that rank agreement under alternative parameter settings and observation subsets should be reported explicitly.
3. Spatial structure should be assessed formally, using diagnostics and maps that can show whether severe congestion forms coherent corridors and local clusters rather than isolated points.

Based on these principles, the present study aims to show that a minimal-data workflow built on passive speed aggregates can produce bottleneck rankings that are robust, interpretable, and operationally useful at the scale of the whole city.

Methods and Materials

1. Data source and study area

The analysis is based on hourly speed data covering the road network of Astana. Each record represents a segment-hour observation and includes the segment identifier, timestamp, hourly median speed, posted speed limit, functional road class, segment length, and probe count. After the initial quality checks, the working dataset contained 6.34 million valid observations for 22,333 unique segments collected over a continuous twenty-day period. This provides sufficient

spatial coverage and temporal depth for city-scale bottleneck analysis while remaining consistent with the practical constraints of short-term operational monitoring.

All spatial layers were converted to the EPSG:3857 coordinate reference system so that distances and areas are expressed in metric units and can be used directly in mapping and spatial analysis. The resulting dataset combines temporal traffic information with segment-level geometric and functional attributes, allowing congestion to be evaluated both as a link-level phenomenon and as part of broader corridor structure.

The overall workflow used in the study is outlined in Figure 1. The main pipeline begins with raw hourly speed aggregates and proceeds through quality filtering, free-flow estimation, link-hour travel-time and delay computation, construction of a length-regularized severity metric, rank-stability assessment, concentration analysis, and final bottleneck ranking. In parallel, the same severity metric is used for spatial diagnostics, while the ranking outputs are used for temporal profiling of the most critical links.

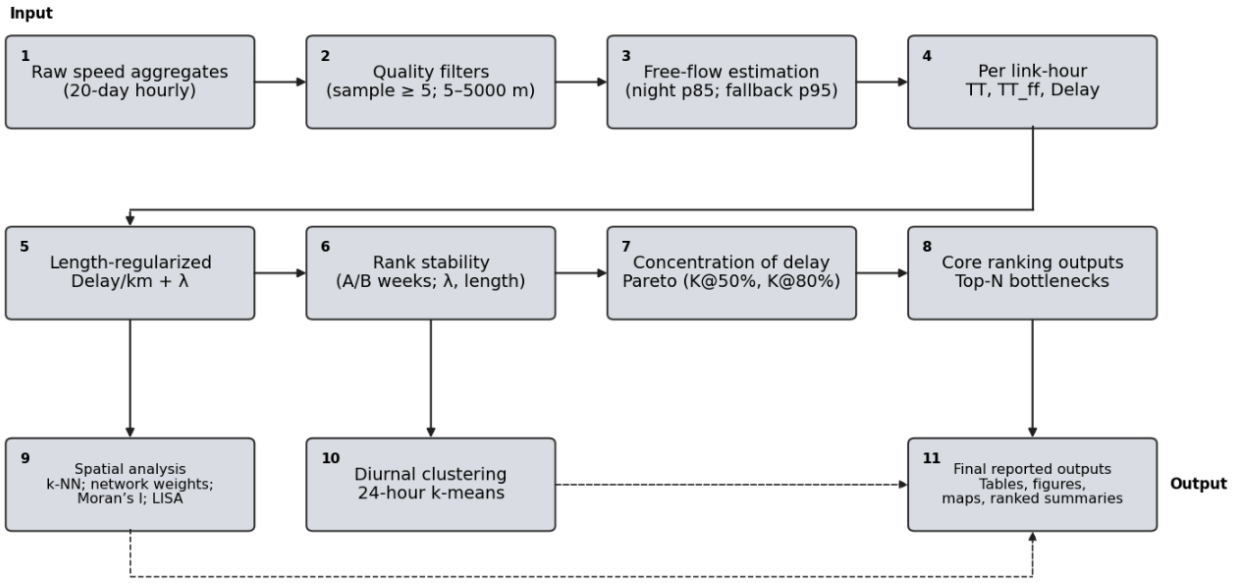

Figure 1. Data pipeline schematic

The key variables used throughout the workflow are summarized in Table 1. Along with the variable names and units, the table clarifies how each attribute is used in filtering, normalization, benchmarking, and later interpretation.

Table 1. Dataset summary

Variable	Description	Unit	Role in analysis
segmentId	Unique road segment identifier	–	Join key for segment-level aggregation, geometry linkage, and final outputs
time	Observation timestamp	–	Temporal index used to construct segment-hour observations and daily profiles
median_kph	Hourly median speed	km/h	Core observed traffic variable used to derive travel time, delay, and free-flow comparisons
speedLimit_kph	Posted speed limit	km/h	Contextual reference used to interpret segment characteristics and assess benchmark plausibility
distance_m	Segment length	m	Used for delay normalization and length-based regularization

frc	Functional road class (1-7)	categorical	Used for stratified descriptive analysis and reliability summaries
sampleSize	Number of probes per hour	–	Used for quality filtering to exclude weakly observed hourly records

2. Data filtering and coverage accounting

The dataset was cleaned before analysis using a small set of explicit filtering rules designed to improve the reliability and comparability of segment-hour observations. The criteria were as follows:

1. Probe count threshold. Only records with sampleSize ≥ 5 were retained. Very small probe counts can make hourly median speed estimates unstable and overly sensitive to isolated observations.
2. Segment length limits. Only links with 5 m ≤ distance ≤ 5 km were retained. Extremely short segments are more vulnerable to geometry artifacts and segmentation noise, whereas very long segments may represent over-aggregated road sections that weaken local bottleneck comparability.
3. Coverage-based selection. Segment coverage was calculated as the share of observed hours within the full 480-hour window, and the joint distribution of coverage and segment length was used to define the subset for further analysis.

Based on these rules, 22,333 total segments were retained after the initial cleaning stage. Coverage quantiles were as follows: 5th percentile = 0.002, 25th = 0.083, 50th = 0.781, 75th = 0.998, and 95th = 1.000. This indicates that half of all retained segments were observed during at least 78% of the study window, confirming dense temporal sampling for a large share of the network.

The results of the filtering stages are summarized in Table 2. The intersection of length ≥ 50 m and coverage ≥ 40% defines the main analysis set containing 8,634 segments. This selection balances spatial representativeness and temporal completeness while limiting distortions from very short or weakly observed links.

Table 2. Retained segment counts under alternative length and coverage thresholds.

Scenario	**Segments**	**Share (%)**
All unique segments	22,333	100.0
Length ≥ 50 m	14,103	63.1
Length ≥ 100 m	7,324	32.8
Length ≥ 200 m	2,272	10.2
Coverage ≥ 20 %	15,475	69.3
Coverage ≥ 40 %	14,087	63.1
Coverage ≥ 60 %	12,764	57.2
Coverage ≥ 80 %	10,955	49.1
Length ≥ 50 m ∩ Coverage ≥ 40 % (analysis set)	**8,634**	**38.7**

As shown in Figure 2, short segments dominate the network and are therefore more likely to introduce instability into normalized congestion measures if no correction is applied. Nearly two-thirds of all links are shorter than 100 meters. This distribution provides a direct empirical rationale for using length-based normalization and mild regularization in the later ranking stage so that very short links do not exert disproportionate influence on the bottleneck list.

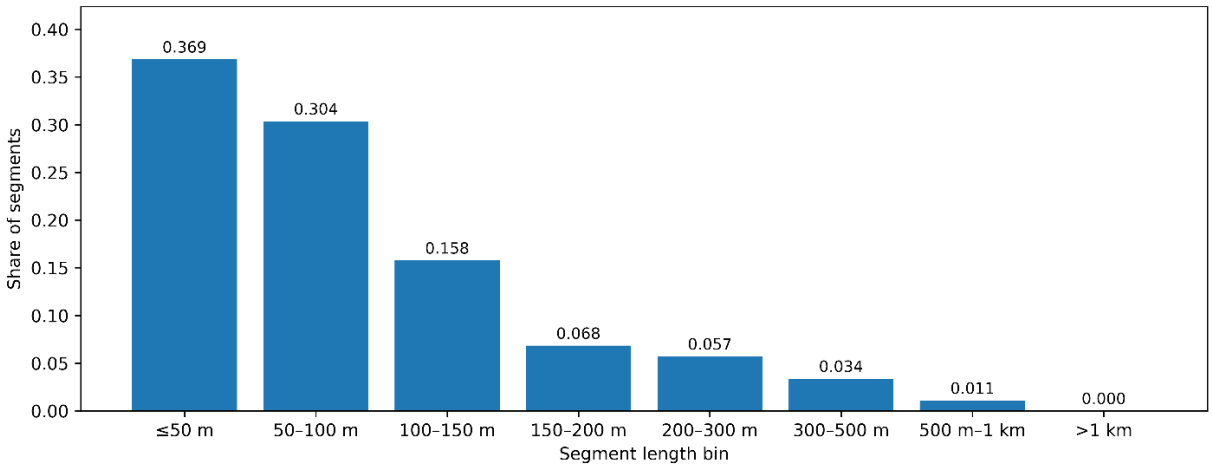

Figure 2. Segment length distribution across the full network

Complementing this, the empirical coverage distribution shown in Figure 3 indicates that most retained segments maintain high hourly availability across the twenty-day period. The steep upper part of the distribution shows that a large share of segments is observed during most of the study window. This supports the use of the 40% coverage threshold for the main analysis set, since it preserves a broad and temporally reliable subset without restricting the analysis to only fully observed links.

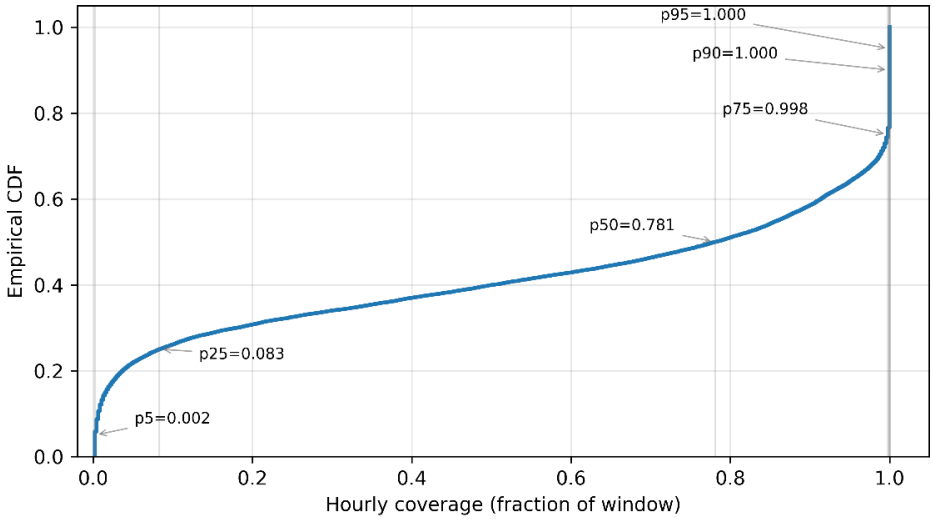

Figure 3. Empirical cumulative distribution of hourly coverage

3. Travel-time, free-flow, and delay computation

Segment-level travel times are derived from observed speeds, and benchmark "free-flow" speeds are used to quantify delay. Travel time for segment *s* at hour *t* is calculated from observed speed as in (1):

$$TT_{s,t} = L_s / v_{s,t} \,, \tag{1}$$

where L_s is segment length (km) and $v_{s,t}$ is hourly median speed (km/h).

Free-flow travel time $TT_{ff}(s)$ was computed using the segment's reference speed $v_{ff}(s)$, as in (2):

$$TT_{ff}(s) = L_s / v_{ff}(s) \,. \tag{2}$$

Hourly delay was defined as the positive difference between actual and free-flow travel times, which prevents negative values under free-flow conditions:

$$\Delta_{s,t} = \max\left(0, TT_{s,t} - TT_{ff}(s)\right). \quad (3)$$

The baseline free-flow speed $v_{ff}(s)$ was defined as the 85th percentile of hourly speeds observed between 01:00 and 05:00. This choice was intended to represent near-uncongested movement while avoiding overly optimistic benchmark values. A night-time window was preferred because it is the most plausible period for approximating free-flow conditions without relying on road-design assumptions alone. At the same time, the benchmark was deliberately kept conservative by using the 85th percentile rather than the maximum observed speed.

To improve reproducibility, a minimum of ten night-time observations was required for the segment-specific night benchmark. When overnight data were insufficient, the 95th percentile of all available observations for that segment was used as a fallback. Because this benchmark choice can influence the ranking, it was not treated as self-evidently correct. Instead, its sensitivity to alternative percentile and time-window definitions was evaluated explicitly later in the Results section.

4. Delay normalization and regularization

To make different road lengths comparable, total delay for each segment was summed over time and divided by distance, giving the value in hours per kilometre:

$$\text{Delay\_per\_km}(s) = \frac{\sum_t \Delta_{s,t}/3600}{L_s}. \quad (4)$$

Short segments tend to vary more because they contain fewer observations. To reduce this noise, a small length-based regularization factor was applied:

$$\text{Delay}_\lambda(s) = \frac{L_s}{L_s+\lambda}\,\text{Delay\_per\_km}(s),\ \lambda > 0\,. \quad (5)$$

In the implementation used for this study, the regularization parameter was set to $\lambda = 0.20$ km.

This formulation dampens fluctuations for short links while leaving longer segments effectively unchanged. The resulting metric remains interpretable and supports stable ranking, concentration analysis, and spatial diagnostics without allowing extremely short urban links to dominate the bottleneck shortlist merely because of volatility.

5. Spatial analysis framework

Spatial relationships in congestion were examined under two alternative weighting schemes. The first linked each segment to its eight nearest neighbors in geographic space. This provided continuity with a standard exploratory spatial approach and served as a geometric baseline.

The second weighting scheme used a network-based adjacency matrix derived from shared segment endpoints. Under this specification, two links were treated as neighbors if they were directly connected in the road network. This representation is closer to actual corridor structure than simple planar proximity and is therefore better suited for checking whether severe congestion forms connected bottleneck corridors.

Global Moran's I was computed under both specifications, and significance was assessed using 999 random permutations. A positive Moran's I indicates that segments with similar severity values tend to group together, while negative values would indicate dispersion.

Local cluster analysis was then carried out under the network-based specification. This made it possible to separate statistically significant corridor cores from adjacent transition links and to determine whether the ranked bottlenecks form connected spatial structures rather than isolated segment-level anomalies.

6. Temporal profiling and clustering

Temporal variation was studied by converting each segment's delay series into normalized 24-hour profiles. This step was applied to the highest-ranked bottlenecks rather than to the whole network, because the goal was to examine the daily behavior of the most critical links.

The resulting profiles were grouped using k-means clustering. The silhouette score was used to select the number of clusters so that the final solution would balance interpretability and within-cluster consistency. This part of the analysis was intended to show whether the most severe bottlenecks follow different daily regimes and whether those regimes differ in their intensity and functional-road-class composition.

Spatial analysis and temporal clustering serve different purposes. The spatial block tests whether bottlenecks form connected corridor structures. The temporal block examines how the most severe bottlenecks evolve over the course of the day. Together, these two views give a more complete picture of recurrent congestion patterns.

7. Robustness checks with external disturbance days

To assess whether the bottleneck ranking is driven by unusual outside conditions, an auxiliary disturbance-day annotation was compiled for the twenty-day study period. Two categories were used:

1. Official extreme-weather days
2. Major citywide disruption days

Weather annotations were derived from the daily weather record linked to the traffic observation window. Disruption days were flagged from event-based annotations covering large road works, security restrictions, and public events with possible network-wide effects.

These labels were not incorporated into the severity metric itself, because the study was designed as a lightweight ranking-first workflow for data-constrained settings rather than as a causal model of congestion. Instead, they were used in a robustness analysis in which the ranking procedure was repeated after excluding the flagged days and then compared with the full-sample baseline using Spearman correlation, Top-100 overlap, and the Pareto thresholds K50 and K80.

This procedure makes it possible to test whether the main bottleneck structure remains visible after removing atypical days. In this way, the study checks whether the ranking reflects recurring congestion patterns rather than a small number of unusual events.

Results

1. Validation of Free-Flow Benchmarks

A reliable free-flow baseline is needed to ensure that the measured delays reflect actual excess travel time rather than bias in the benchmark itself. Using the benchmark defined in the previous section, the estimated free-flow speeds were compared with the posted speed limits for all 22,333 segments.

The comparison with posted speed limits is given in Figure 4. Most points are concentrated close to the 1:1 line, which indicates good agreement between the estimated free-flow speeds and the official limits. The main deviations appear on short local segments, where limited probe coverage can lead to greater variation in percentile-based estimates. Overall, this pattern supports the plausibility of the benchmark and shows that it provides a reasonable reference for delay estimation across the network.

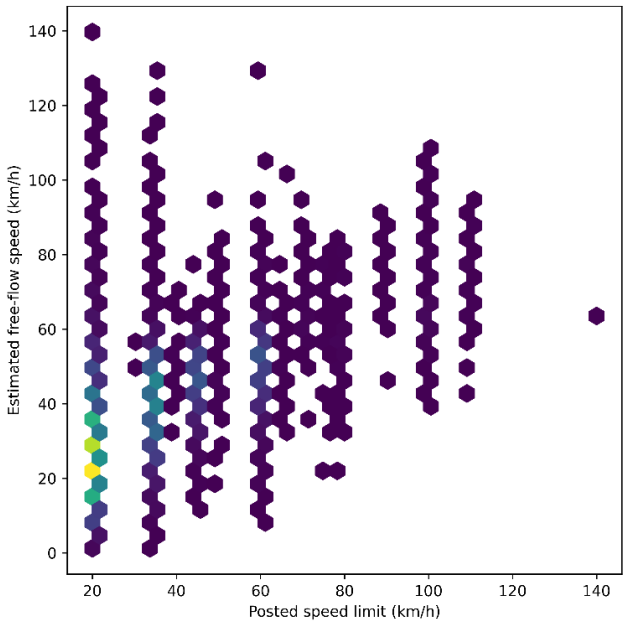

Figure 4. Comparison of estimated free-flow speeds and posted speed limits

The benchmark was also checked under alternative free-flow definitions to determine whether the final ranking depends strongly on one specific specification. Separate sensitivity tests were carried out for six alternative benchmark scenarios obtained by varying either the percentile threshold or the definition of the night-time window. The resulting rank agreement is reported in Table 3.

Table 3. Sensitivity of bottleneck ranking to alternative free-flow definitions

Scenario	**Spearman vs baseline**	**Top-100 overlap**	**K50**	**K80**
80th percentile, 01:00–05:00	0.997	94	1312	3797
85th percentile, 00:00–04:00	0.990	95	1298	3779
85th percentile, 01:00–05:00 (baseline)	1.000	100	1336	3868
85th percentile, 02:00–06:00	0.990	98	1348	3897
90th percentile, 01:00–05:00	0.996	96	1364	3943
95th percentile, 01:00–05:00	0.985	90	1408	4037

Across all tested scenarios, the bottleneck ranking remained highly consistent with the baseline. Spearman correlation ranged from 0.985 to 1.000, and Top-100 overlap ranged from 90 to 100 segments. The concentration thresholds changed only slightly, with K50 ranging from 1,298 to 1,408 and K80 from 3,779 to 4,037. Overall, the results indicate that the main bottleneck structure remains stable under several reasonable definitions of free-flow speed.

2. Network-Wide Delay Distribution

After establishing the free-flow baseline, normalized congestion severity was calculated for the 8,634 valid segments in the main analysis set. The resulting distribution of delay per kilometre is strongly right-skewed. Most segments show low delay, while a much smaller share displays very high values.

This pattern is illustrated in Figure 5. Short links exhibit higher variability, as small changes in traffic conditions produce larger shifts in normalized delay. By contrast, medium and long segments show lower and more consistent delay levels. This difference confirms the need to account for segment length when comparing congestion across the network.

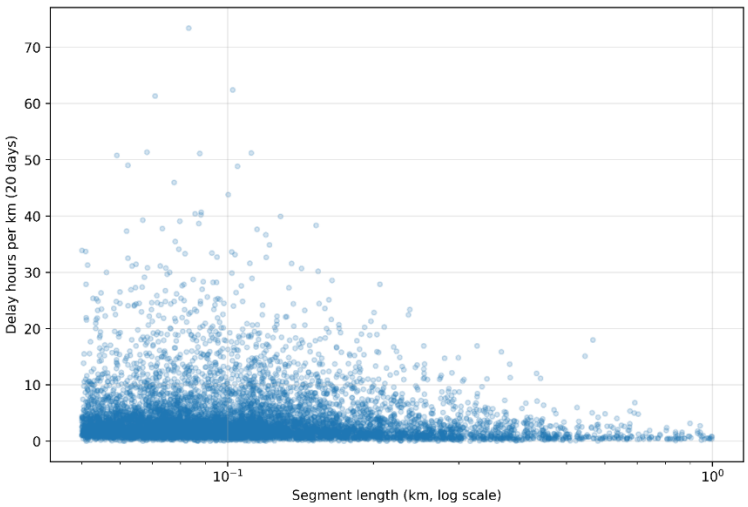

Figure 5. Relationship between delay hours per km and segment length

Together, these results support two key choices in the metric design. First, congestion severity is expressed in hours per kilometre so that links of different lengths can be compared more fairly. Second, length-based regularization is used to reduce instability on short links without materially changing the position of longer ones. In this form, the severity indicator remains both interpretable and suitable for ranking and clustering.

3. Concentration of Cumulative Delay

To examine how delay is distributed across the city network, cumulative delay was ranked by length-adjusted severity. The resulting concentration curve is shown in Figure 6.

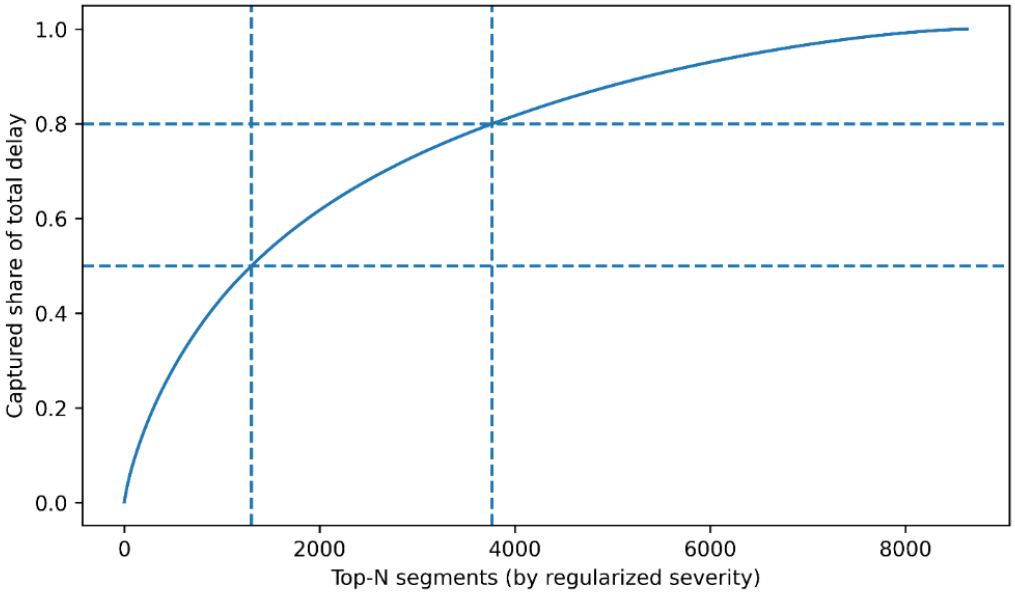

Figure 6. Pareto curve of cumulative delay share

The concentration pattern is strong. The top 1,336 segments, or about 15.5% of the analysis set, account for half of the total delay. The top 3,868 segments, or about 44.8%, account for 80% of the total delay. This means that a relatively limited share of the network carries a disproportionately large burden of congestion. In practical terms, the result suggests that targeting a manageable subset of recurring bottlenecks could yield substantial network-wide gains.

Ranking robustness was also examined under alternative values of the regularization parameter λ. The corresponding Spearman correlation and Top-100 overlap are presented in Figure 7.

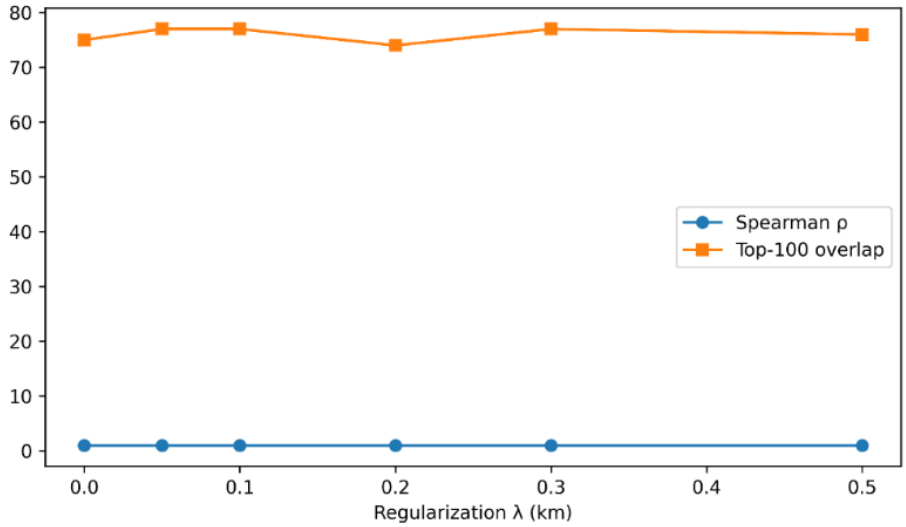

Figure 7. Rank stability across regularization λ (Spearman ρ and Top 100 overlap)

The ranking remained stable across the tested values. Correlation values stayed consistently high, and Top-100 membership changed only slightly. This indicates that the identified bottlenecks are not a product of one narrow parameter choice but reflect a stable congestion structure in the data.

4. Robustness to External Disturbance Days

Additional robustness checks were carried out to determine whether the ranking is overly influenced by unusual outside conditions. The ranking procedure was repeated after excluding days with official extreme-weather conditions and after excluding days marked by major citywide disruptions. The results are summarized in Table 4.

Table 4. Robustness of the bottleneck ranking under exclusion of external-disturbance days

Analysis scenario	**Days**	**Spearman vs baseline**	**Top-100 overlap**	**K50**	**K80**
Full sample	20	1.000	100	1336	3868
Excluding official extreme-weather days	18	0.994	95	1330	3857
Excluding major disruption days	9	0.980	86	1296	3794
Excluding weather and major disruption days	9	0.980	86	1296	3794

The ranking remained stable in all subsets. Excluding official extreme-weather days produced a Spearman correlation of 0.994 and retained 95 of the baseline Top-100 links, while excluding major disruption days produced a correlation of 0.980 and retained 86 Top-100 links. The concentration thresholds changed only modestly, with K50 ranging from 1,296 to 1,330 and K80 from 3,794 to 3,857. This indicates that the main bottleneck structure is not driven by a small number of atypical days.

In the present sample, the official extreme-weather days fall within the broader set of major disruption days. As a result, excluding both categories produces the same retained-date subset as excluding major disruption days alone, which explains the identical values in the last two rows of Table 4.

A descriptive comparison of disturbance regimes leads to the same conclusion. Major disruption days alone remain close to normal conditions, whereas the combined weather-and-disruption regime shows higher delay levels, lower relative speeds, and a larger congested share. The same pattern is even more pronounced within the Top-100 bottleneck subset. This indicates that adverse external conditions amplify already vulnerable links, but the core bottleneck structure itself remains persistent rather than being created by a small number of abnormal episodes.

5. Identification of Major Bottlenecks

The ten highest-severity segments, ranked by length-regularized delay, are reported in Table 5. Where source metadata were available, street or corridor names are provided to make the ranking easier to interpret in network terms.

Table 5. Top ten bottlenecks by length-regularized delay

Rank	Street / corridor	FRC	Mean delay (h/km)	Regularized severity (Delay_shr)	Cumulative share (%)
1	Uly Dala Avenue	4	3.058	21.552	14.73
2	Metadata unavailable	7	2.600	21.145	27.25
3	Kaiyma Mukhamedkhanova Street	3	2.134	18.373	37.53
4	Metadata unavailable	6	2.035	16.793	47.33
5	Kosmonavtov Street	6	1.597	16.573	55.03
6	Uly Dala Avenue	4	2.555	16.041	67.34
7	Saken Seifullin Street	6	1.664	15.634	75.35
8	Akhmeta Zhubanova Street	6	2.130	15.572	85.61
9	Kurgalzhyn Highway	5	1.825	14.633	94.41
10	Metadata unavailable	7	1.161	14.148	100.00

The top bottlenecks are concentrated along mid-tier arterials and collector roads characterized by sustained queuing near signalized intersections and merge points. Several named links recur among the highest-ranked segments, including Uly Dala Avenue, Kaiyma Mukhamedkhanova Street, Kosmonavtov Street, Saken Seifullin Street, Akhmeta Zhubanova Street, and Kurgalzhyn Highway. A few top-ranked links still remain without street metadata in the merged source layer, but this affects only the naming and not the ranking itself.

Across the top ten, mean delay per kilometre ranges from 1.161 to 3.058 hours/km. Functionally, the list is dominated by collectors and minor arterials, especially FRC 6 together with FRC 4 and FRC 7. This composition suggests that a large share of citywide delay is generated at the collector-arterial interface, where repeated queuing and short-cycle inefficiencies can accumulate over time.

The spatial context of these high-ranked links is given in Figure 8. The mapped pattern shows that severe bottlenecks are not spread evenly across Astana, but are concentrated within a limited number of recurrent corridors. Panel (b) highlights the central core, while panels (c) and (d) show recurring southern and eastern corridor systems where the most severe links form continuous arterial structures rather than isolated segment-level anomalies.

Figure 8. Regularized bottleneck severity across the analysed network.

Taken together, the ranked shortlist and the mapped critical segments point to a corridor-based congestion structure rather than isolated local anomalies. This makes the ranking more useful for practical intervention planning, since it identifies recurring corridors where targeted measures are more likely to have system-level effects.

6. Spatial Autocorrelation of Delay Severity

Spatial clustering of congestion severity was then evaluated under the two weighting schemes described in the Methods section. The comparison of global Moran's I values is reported in Table 6. The network-based specification produced the stronger positive autocorrelation, indicating that congestion severity is organized more clearly along connected road corridors than under the purely geometric baseline.

Table 6. Comparison of global Moran's I under alternative spatial-weight specifications

Spatial weights	Moran's I	p-value
8-nearest neighbors	0.213931	0.001
Network endpoints	0.251760	0.001

Both specifications yielded statistically significant positive Moran's I values, which confirms that severe delay is spatially structured rather than random. Under the 8-nearest-neighbour specification, Moran's I equals 0.213931. Under the network-endpoint specification, the value rises to 0.251760. The stronger coefficient under network connectivity indicates that corridor structure explains the spatial organization of congestion at least as strongly as simple planar proximity.

Local cluster analysis was then carried out under the network-based specification. The significant local cluster types are presented in Figure 9. Panel (a) gives the whole-network pattern, while panel (b) shows the central zoom where High-High corridor cores and adjacent Low-High transition links can be seen most clearly.

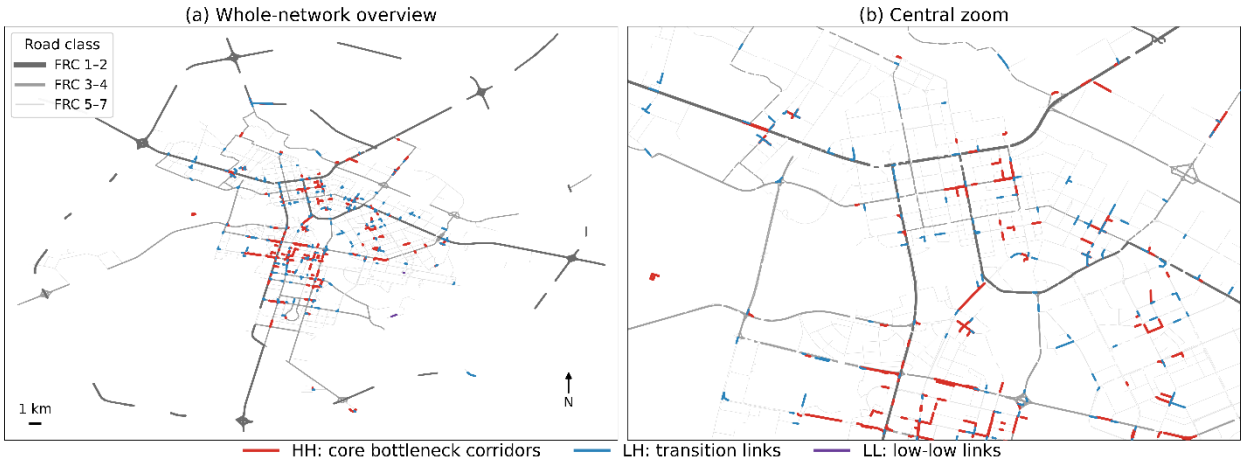

Figure 9. Network-based local spatial clustering of bottleneck severity.

High-High segments represent the core bottleneck corridors, where severe links are surrounded by similarly severe neighbours. Low-High segments identify adjacent transition links whose own severity is lower but which are directly connected to highly congested links. Low-Low segments are almost absent in the significant results, and no High-Low links were detected. This configuration is consistent with a corridor-style congestion structure, in which the strongest local signal comes from recurrent bottleneck cores and their immediate connected surroundings rather than from large uniform low-delay zones.

The numerical distribution of local cluster types is provided in Table 7.

Table 7. Counts of local cluster types under network-based permutation testing.

Quadrant	Significant = True	Significant = False	Total
HH	293	1060	1353
HL	0	1001	1001
LH	225	1311	1536
LL	2	4339	4341
Total	520	7711	8231

Of the 8,231 evaluated links with valid network neighbours, 520 were statistically significant, which is about 6.3% of the local network-based analysis set. Among these, 293 links were classified as High-High and 225 as Low-High. No High-Low links and only 2 Low-Low links were significant. This pattern shows that the strongest local spatial signal comes from core bottleneck corridors and their adjacent transition links rather than from large homogeneous low-delay zones.

Taken together, the global and local results show that congestion in the study network is not only concentrated in the ranking sense, but also organized spatially in connected corridor structures. This supports corridor-level diagnosis and treatment rather than isolated intervention at single segments.

7. Diurnal Clustering of Congestion Profiles

To examine temporal variation, normalized 24-hour delay vectors were constructed for the 500 highest-ranked bottleneck segments. These profiles were then clustered using k-means in order to identify recurring daily delay regimes within the most critical part of the network.

Silhouette diagnostics indicated that three clusters provide the best balance between interpretability and within-cluster consistency for the top-500 set. The resulting centroids are presented in Figure 10.

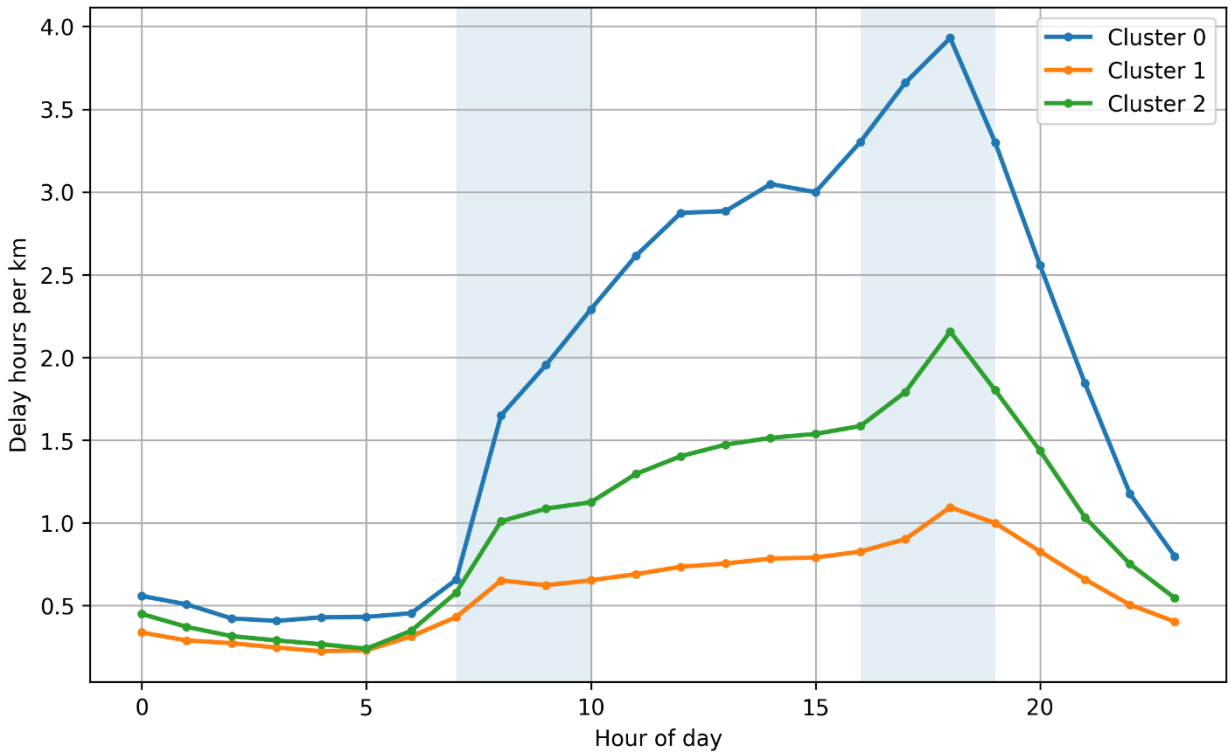

Figure 10. Cluster centroids of bottleneck delay profiles (k = 3, top-500).

All three clusters display a common evening-oriented structure, with the highest average delay occurring at 18:00. The main difference between the clusters is therefore not the timing of the daily maximum, but the intensity of the evening buildup. Cluster 0 represents the most severe profile and contains 24 segments, or 4.8% of the top-500 set. Cluster 2 represents an intermediate regime with 151 segments, or 30.2%. Cluster 1 is the largest group, with 325 segments, or 65.0%, and represents the mildest profile. Overall, the temporal structure of the most critical bottlenecks is dominated by a shared evening peak, while the clusters mainly distinguish between severe, moderate, and mild delay regimes.

The functional composition of these three clusters is shown in Figure 11 and reported numerically in Table 8.

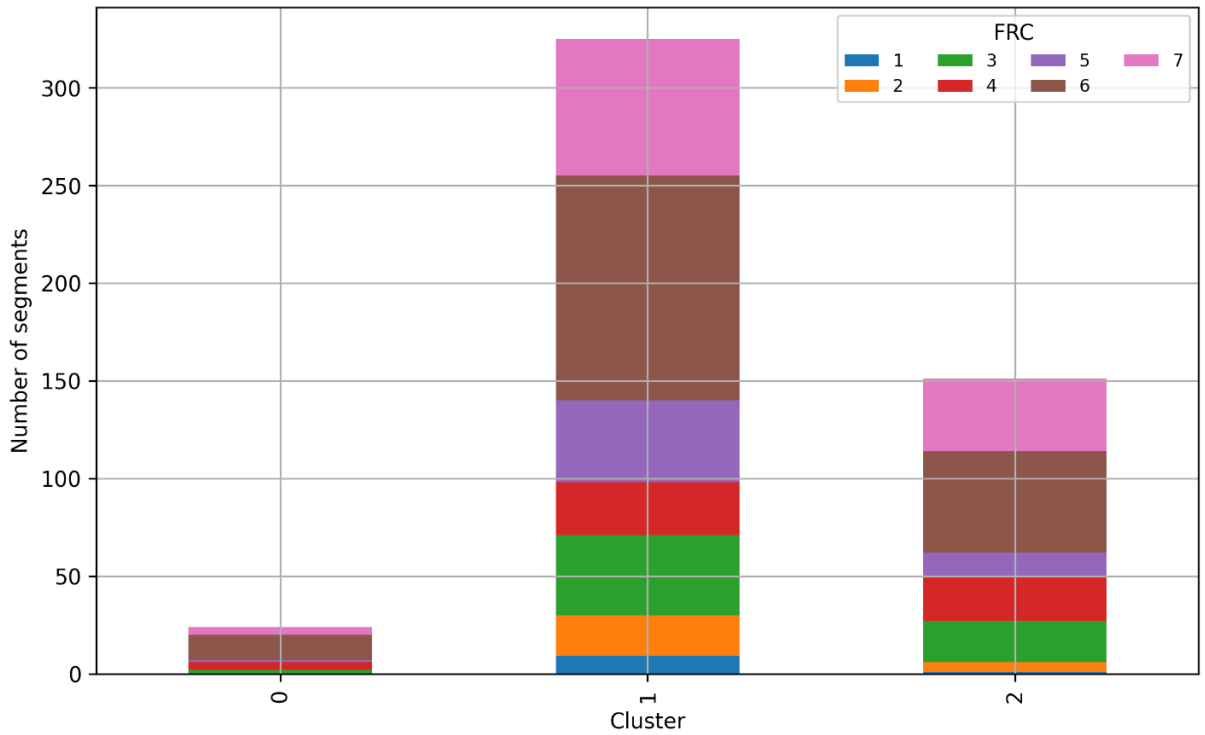

Figure 11. Cluster composition by FRC (k = 3, top-500).

Across all three groups, the largest shares belong to collector and lower-tier urban roads, especially FRC 6 and FRC 7, with additional contributions from FRC 3, FRC 4, and FRC 5. Cluster 0, although small, is concentrated mainly in FRC 6 and FRC 7 links, with a smaller contribution from FRC 4. Cluster 1, the dominant group, is also heavily represented by FRC 6 and FRC 7, but includes substantial counts from FRC 3, FRC 4, and FRC 5. Cluster 2 occupies an intermediate position and remains dominated by FRC 6 and FRC 7, while also including visible contributions from FRC 3 and FRC 4.

Table 8. Distribution of diurnal clusters by FRC.

FRC	Cluster 0	Cluster 1	Cluster 2	Total
1	0	9	1	10
2	0	21	5	26
3	2	41	21	64
4	4	27	23	54
5	1	42	12	55
6	13	115	52	180
7	4	70	37	111
Total	24	325	151	500

This distribution suggests that the diurnal regimes are shaped less by completely different road classes than by different levels of congestion intensity within broadly similar urban corridor types. From an operational point of view, the results point mainly to the evening period as the critical time for traffic management, while also suggesting that not all bottlenecks require the same treatment intensity.

Discussion

This study examined whether passive speed aggregates, combined with transparent filtering and normalization steps, can support practical identification of urban bottlenecks. The results show that they can. The free-flow benchmark was close to posted speed limits across the

network, which suggests that the estimated delays reflect actual operational losses rather than artefacts of the method. Sensitivity checks also showed that the main ranking remains stable under alternative free-flow definitions, which strengthens confidence in the resulting bottleneck structure.

Delay was strongly concentrated on a limited subset of segments even after regularization and robustness checks. In practical terms, this means that traffic managers can prioritize a smaller set of recurring problem corridors rather than dispersing attention across the whole network.

The spatial results reinforce this interpretation. Figure 8 and Table 5 show that the highest-severity links are concentrated within a limited number of recurrent corridors, while Figure 9 indicates that these links form connected local structures rather than isolated hot spots. In practical terms, this supports corridor-scale measures such as signal retiming, turn-bay adjustments, and local capacity management.

The comparison between geometric and network-based spatial weights adds an important methodological point. Positive spatial autocorrelation remained significant under both specifications, but the network-based matrix produced stronger clustering. This suggests that congestion in the study area is better understood through actual road connectivity than through geographic proximity alone. The local results point mainly to connected High-High corridor cores and adjacent transition links. This pattern supports a corridor-based view of congestion and is consistent with how recurrent queuing is usually observed in urban practice.

Temporal clustering adds a complementary perspective. All three groups share a common evening-oriented structure, and the main difference between them lies in intensity rather than peak timing. This suggests that the highest-ranked bottlenecks follow similar daily rhythms but differ in the severity of late-day buildup.

Two design choices were especially important. Normalizing severity by distance made comparisons fair across links with very different lengths. Mild regularization reduced volatility on short segments without changing the relative position of longer ones. Together these steps improved interpretability and limited the influence of random variation. Because all calculations rely on simple and reproducible inputs, the framework is suitable for routine monitoring and for integration with digital twin platforms.

Several limitations should still be acknowledged. The twenty-day window does not capture seasonal patterns or rare special events. Night-time percentiles were used as a proxy for free-flow conditions, and limited overnight data may affect the robustness of this benchmark on some links. Although the study compares geometric and network-based spatial weights, the local spatial model still simplifies actual propagation because it does not encode turning movements, lane-level connectivity, or signal-control logic. The results also depend on the quality of the probe data and on its uneven distribution across roads and time.

Even with these limitations, the findings show that large-scale passive speed data can bridge big-data analytics and day-to-day traffic management when they are processed in a disciplined and transparent way. The method makes it possible to identify key corridors, match treatments to daily traffic rhythms, and track performance with indicators that can be reproduced. Future extensions can build on this base by using longer observation windows, richer external annotations, and more detailed network representations.

Conclusion

This study shows that large-scale passive speed data can serve as a dependable basis for congestion diagnostics within a digital twin of an urban road network. Using percentile-based free-flow benchmarks, length-normalized delays, and simple spatial and temporal analysis, the workflow identifies recurring bottlenecks in a consistent and interpretable way. It does so without relying on dedicated sensors or complex model calibration.

The results show that congestion in the study area follows a clear and stable pattern. Total delay is concentrated on a limited set of corridors, which suggests that targeted improvements on a relatively small group of critical links could generate substantial network-wide gains. The ranking

also remained stable under alternative benchmark definitions, regularization settings, and disturbance-day exclusions, which supports the robustness of the identified bottleneck structure.

Spatial and temporal analyses further clarified the bottleneck pattern. Severe delay is organized in connected corridor structures, the network-based specification captures this structure more clearly than the geometric baseline, and the highest-ranked bottlenecks share a common evening-oriented buildup that differs mainly in intensity.

The main methodological contribution of the study lies in its transparency, scalability, and modest data requirements. In practical terms, the framework turns routinely collected probe data into interpretable indicators that support corridor prioritization and evidence-based traffic management within a digital twin environment.

Future work should extend the observation period to capture seasonal variation and rare events. It should also validate the findings using independent data sources and explore how macroscopic delay patterns can be linked to more detailed network representations within a digital twin framework. Overall, the study shows that transparent and reproducible methods built on passive speed aggregates can help cities move from reactive traffic control toward more predictive and network-oriented management.

References

[1] N. Serok, S. Havlin, and E. B. Lieberthal, “Identification, Cost Evaluation, and Prioritization of Urban Traffic Congestions and Their Origin”, *Scientific Reports*, vol. 12, no. 13026, 2022, doi: 10.1038/s41598-022-17404-8.

[2] L. Wei, P. Chen, Y. Mei, and Y. Wang, “Turn-level network traffic bottleneck identification using vehicle trajectory data”, *Transportation Research Part C: Emerging Technologies*, vol. 140, 103707, 2022, doi: 10.1016/j.trc.2022.103707.

[3] J. Duan, G. Zeng, N. Serok, D. Li, E. B. Lieberthal, H.-J. Huang, and S. Havlin, “Spatiotemporal Dynamics of Traffic Bottlenecks Yields an Early Signal of Heavy Congestions”, *Nature Communications*, vol. 14, 2023, doi: 10.1038/s41467-023-43591-7.

[4] A. Cottam, X. Li, X. Ma, and Y.-J. Wu, “Large-Scale Freeway Traffic Flow Estimation Using Crowdsourced Data: A Case Study in Arizona”, *Journal of Transportation Engineering, Part A: Systems*, vol. 150, no. 7, 04024030, 2024, doi: 10.1061/JTEPBS.TEENG-8304.

[5] Q. Cheng, Z. Liu, J. Lu, G. List, P. Liu, and X. S. Zhou, “Using Frequency Domain Analysis to Elucidate Travel Time Reliability Along Congested Freeway Corridors”, *Transportation Research Part B: Methodological*, vol. 184, 102961, 2024, doi: 10.1016/j.trb.2024.102961.

[6] J. Zang, P. Jiao, S. Liu, X. Zhang, G. Song, and L. Yu, “Identifying Traffic Congestion Patterns of Urban Road Network Based on Traffic Performance Index”, *Sustainability*, vol. 15, no. 2, 948, 2023, doi: 10.3390/su15020948.

[7] Z. Hou, H. M. Zhang, and X. Zhang, “A Comparative Study of Freeway Travel Times from Four Data Providers”, *Transportation Research Record*, 2023, doi: 10.1177/03611981221144290.

[8] Z. Zhang, L.D. Han, Y. Liu, “Exploration and evaluation of crowdsourced probe-based Waze traffic speed”, *Transportation Letters*, vol. 14, no. 8, pp. 797-808, 2022, doi: 10.1080/19427867.2021.1906477.

[9] W. Cheng, J.-L. Li, H.-C. Xiao, and L.-N. Ji, “Combination predicting model of traffic congestion index in weekdays based on LightGBM-GRU”, *Scientific Reports*, vol. 12, 2912, 2022, doi: 10.1038/s41598-022-06975-1.

[10] B. Qiu, W. Fan, “Travel time forecasting on a freeway corridor: a dynamic information fusion model based on the random forests approach”, *Smart and Resilient Transportation*, vol. 3, no. 2, pp. 131-148, 2021, doi: 10.1108/SRT-11-2020-0027.

[11] K.A. Salvi, M. Kumar, and A. M. Hainen, “Sensitivity of Traffic Speed to Rainfall”, *Weather, Climate, and Society*, vol. 14, no. 4, pp. 1165-1175, 2022, doi: 10.1175/WCAS-D-22-0024.1.

[12] D.W. Soole, S. O’Hern, M. Cameron, S. Peiris, S. Newstead, W. Anderson and T. Smith, “Using GPS Probe Speed Data to Estimate the Attribution of Speeding on Casualty Crashes: A Case Study in Queensland”, *Journal of Road Safety*, vol. 34, no. 1, pp. 39-48 2024, doi: 10.33492/JRS-D-22-00003.

[13] T. Veihelmann, V. Shatov, M. Lübke, N. Franchi “Using Probe Counts to Provide High-Resolution Detector Data for a Microscopic Traffic Simulation”, *Vehicles*, vol. 6, no. 2, pp. 747-764, 2024, doi: 10.3390/vehicles6020035.

[14] J. Kim, W. Jeon, and S. Kim, “Analyzing the Relationship Between User Feedback and Traffic Accidents Through Crowdsourced Data”, *Sustainability*, vol. 16, no. 22, 9867, 2024, doi: 10.3390/su16229867.

[15] B. Metzger, A. Loder, L. Kessler, K. Bogenberger, “Spatio-temporal prediction of freeway congestion patterns using discrete choice methods”, *EURO Journal on Transportation and Logistics*, vol. 13, 100144, 2024, doi: 10.1016/j.ejtl.2024.100144.

[16] C.-C. Lin, M.-C. Ho, C.-C. Hung, H.-H. Hsu, “A comparative study and simple baseline for travel time prediction”, *Scientific Reports*, vol. 15, no. 25609, 2025, doi: 10.1038/s41598-025-02303-5.

[17] A. Kandiri, R. Ghiasi, M. Nogal, and R. Teixeira, “Travel time prediction for an intelligent transportation system based on a data-driven feature selection method considering temporal correlation”, *Transportation Engineering*, vol. 18, 100272, 2024, doi: 10.1016/j.treng.2024.100272.

[18] E. B. Lieberthal, N. Serok, J. Duan, G. Zeng, and S. Havlin, “Addressing the urban congestion challenge based on traffic bottlenecks”, *Philosophical Transactions of the Royal Society A*, vol. 382, no. 2285, 2024, doi: 10.1098/rsta.2024.0095.

[19] Z. Jin, Y. Wang, X. Meng, D. Duan, and N. Wang, “Identifying critical roads in urban road networks considering congestion propagation”, *Physica A: Statistical Mechanics and its Applications*, 2025, doi: 10.1016/j.physa.2025.130795.

[20] J. Zhang, Z. Zhang, D. Song, Z. Huang, and L. Lu, “A Study on a Framework for Identifying Critical Roads in Urban Road Traffic Networks Based on the Resilience Perspective Against the Background of Sustainable Development”, *Applied Sciences*, vol. 15, no. 7, 3581, 2025, doi: 10.3390/app15073581.