Incremental Predictive Value of Spatial Graph Structure in Hourly Istanbul Traffic Forecasting: A Confirmatory Multi-Season Study

Faramarz Kowsari

Software Engineer and AI Researcher, Istanbul, Türkiye

ORCID: 0000-0003-1692-0453

Correspondence: https://faramarzkowsari.github.io/

Manuscript version: v1.4
Date: 16 August 2026
Study registration: OSF Registries, DOI: 10.17605/OSF.IO/FM5R7
Versioned project archive: Zenodo, DOI: 10.5281/zenodo.21916357
Code repository: https://github.com/FaramarzKowsari/istanbul-graphtraffic-ai
Research website: https://faramarzkowsari.github.io/istanbul-graphtraffic-ai/

Preprint-ready manuscript for engrXiv

Abstract

I had long assumed that when forecasting traffic in a highly connected and complex city such as Istanbul, providing a model with the physical road network and the actual relationships between streets should offer a substantial predictive advantage. Yet one question kept returning to me: if a model already observes the previous 24 hours of traffic, the time of day, and the day of the week, might it already be learning much of this spatial information indirectly through temporal patterns? To avoid the temptation of unconsciously modifying the analysis or selecting a more favorable result, I constrained the study before examining the confirmatory outcomes. The main research question, confirmatory months, model conditions, random seeds, statistical procedures, and primary evaluation metric were specified in advance, so that I could not later change the data or models simply to produce a more attractive result. Using hourly traffic data from Istanbul Metropolitan Municipality (İBB) for May and November 2024, I studied 64 traffic locations per analyzable month and compared a Dynamic Graph Transformer (DGT) supplied with physical road-network structure against a temporal MLP that used no spatial adjacency information. Both models used the preceding 24 hours as input. At the +1 hour horizon, the graph model achieved a mean MAE of 3.6804 km/h compared with 3.6986 km/h for the temporal MLP. This corresponds to a paired difference of -0.0181 km/h, or an estimated improvement of approximately 0.49%. The difference was not statistically reliable: the preregistered primary test yielded $p = 0.5000$, and the 95% hierarchical bootstrap confidence interval [-0.1760, +0.1120] km/h crossed zero. A secondary +1 hour versus +6 hour contrast produced a raw p value of 0.0137, but the Holm-adjusted p value was 0.0684 and therefore did not meet the registered significance criterion. The conclusion is not that road graphs are useless for traffic forecasting. Rather, under this specific experimental design, these two confirmatory months, and the tested architecture, adding explicit physical road structure did not provide a reliable predictive advantage over a strong temporal model. One possible explanation is that temporal regularities or adaptive relationships learned by the graph model already capture part of the information represented by the physical network. The result illustrates how an engineering idea can appear compelling and intuitively correct while real data show that its incremental contribution is smaller than expected.

Keywords: traffic forecasting; graph neural networks; Dynamic Graph Transformer; spatiotemporal learning; Istanbul; intelligent transportation systems; preregistration; reproducible research; road graphs; adaptive adjacency

1. Introduction

I have never thought of Istanbul as simply a dense and congested point on a map. To me, it is a living system: a city where two continents meet and where bridges, highways, arterial roads, and intricate local streets form a deeply interconnected network. Forecasting traffic in such a city is therefore not merely a matter of asking how fast vehicles are moving at a particular moment. The more interesting problem is understanding how conditions at one location propagate through the network and how disruption or congestion at one node may influence other parts of the city. This chain-like behavior of urban mobility is what first made the problem scientifically compelling to me.

In trying to predict such a system, we usually provide a model with two broad kinds of information. The first is temporal. A model can observe the previous 24 hours, learn the time of day and day of the week, and absorb recurring patterns such as morning acceleration, evening congestion, and daily or weekly traffic rhythms. The second is spatial. A model can be told which locations are physically connected, in which direction vehicles can travel, how long travel between locations takes, and how the road network is actually structured. This led me to a simple but important question: if temporal information is already rich enough for a model to learn much of the city's recurring behavior, does explicitly providing the physical road network still contribute substantial new predictive information?

A good result from a graph neural network does not, by itself, answer that question. It is tempting to observe strong performance from a graph-based model and conclude that the physical graph must have been responsible. I was not satisfied with that interpretation. The graph model may simply have a stronger architecture. Its adaptive adjacency mechanism may already discover useful relationships among locations. The temporal history may carry most of the predictive signal, leaving the physical road graph with only a marginal role. My question was therefore not whether the Dynamic Graph Transformer was a capable model. The stricter question was whether adding explicit, directed physical road structure created measurable predictive value beyond what could already be learned from temporal information and adaptive relationships.

Machine-learning experiments also offer researchers an unusually large number of choices. If a result is disappointing, it is easy to try another month, another random seed, another set of nodes, another graph configuration, or another forecast horizon until a more attractive result appears. Such flexibility can influence conclusions even without deliberate manipulation. I therefore wanted to reduce my own freedom before looking at the confirmatory outcomes. The main research question, confirmatory months, model conditions, random seeds, evaluation criteria, and statistical analysis plan were specified in advance. My intention was simple: once the confirmatory analysis began, the data should determine the answer rather than my preference for a particular result.

Within this frozen framework, I analyzed hourly traffic data from Istanbul Metropolitan Municipality (İBB) for May and November 2024 across 64 traffic locations per analyzable month. Both model families received a 24-hour traffic history. On one side was a temporal MLP with no spatial adjacency information. On the other was a Dynamic Graph Transformer supplied with a directed road-travel graph together with a learned adaptive adjacency component. Forecasts were evaluated at +1, +2, +3, and +6 hours, with the +1 hour horizon defined in advance as the primary confirmatory endpoint.

The purpose of this study is therefore not to claim a new benchmark record or to promote architectural complexity for its own sake. It is a focused attempt to isolate and test the incremental predictive contribution of explicit physical road structure under controlled conditions. Because the analysis plan was fixed in advance, the result is reported in the same way whether it strongly supports the original expectation or produces a null finding. For me, the more important question is not whether one model improves MAE by a few hundredths of a unit, but whether the additional complexity we introduce into our models genuinely exposes them to new information, or merely creates an elegant engineering structure whose practical contribution is smaller than intuition suggests.

2. Related Work

2.1 Graph-based traffic forecasting

When we look at the evolution of traffic forecasting, early graph-based models delivered a clear message: traffic should not be treated simply as a collection of independent and unrelated time series. Sensors and measurement locations exist within a road network, and conditions at one point can influence conditions elsewhere. DCRNN used directed graphs and diffusion processes to model spatiotemporal traffic dynamics [1]. STGCN combined graph convolution with temporal convolution [2], while ASTGCN introduced spatial and temporal attention mechanisms [3]. These studies reinforced my view that space and time are deeply intertwined in traffic forecasting. Yet one question remained: how much of the predictive improvement in these models actually comes from the physical structure of the road network, and how much comes from the expressive power of the architecture itself?

This uncertainty became stronger with Graph WaveNet and its use of adaptive adjacency [4]. A model no longer had to accept every relationship directly from a physical graph defined by the researcher. It could discover and learn hidden dependencies from the data itself. GMAN [5] and later Transformer-based approaches, building on self-attention [6], extended this direction by modeling spatial and temporal relationships through attention mechanisms. If a model is already capable of learning its own graph relationships from the data, what additional predictive information does an explicitly supplied physical road map provide? This question became one of the main motivations for the present study.

The development of traffic forecasting models has continued with increasingly sophisticated architectures. More recent approaches such as DST-GTN [7], STGormer [9], STGformer [10], GDGCRN [11], and GAPSTGCN [12] have explored dynamic spatial relationships, spatiotemporal heterogeneity, computational efficiency, representation learning, signal decoupling, and pre-training. Each generation has attempted to introduce another aspect of real traffic complexity into the modeling process. Increasing architectural complexity, however, does not necessarily answer the question addressed here. A more complex model may produce lower forecasting error without demonstrating that the physical road graph itself is responsible for the improvement.

2.2 Istanbul-specific traffic graph modeling

When I began focusing specifically on Istanbul traffic data, the work of Olug et al. (2024) was particularly relevant [8]. Their IBB Traffic Graph Data study examined observations from 2,451 Istanbul traffic locations and combined temporal feature engineering, GLEE embeddings, and ExtraTrees for road traffic prediction. The study was important to my work because it demonstrated that İBB traffic data can support meaningful graph-based modeling and benchmarking at a large urban scale. However, their research question was different from mine. My objective was not simply to demonstrate once again that graph representations can be useful for traffic forecasting. I wanted to determine more specifically how much additional predictive value an explicit directed physical road graph contributes when it is placed alongside adaptive adjacency and a strong temporal baseline.

2.3 Research gap and confirmatory ablation

Taken together, the literature reveals the research gap that motivated this study. Previous work shows that graph models can be powerful. Adaptive graph learning demonstrates that models can discover useful relationships without relying entirely on a predefined physical network. Istanbul-specific research has also shown that graph representations are practical for İBB traffic data. Yet an important question remains: does explicit directed physical road structure itself provide incremental predictive value beyond architectural capacity, temporal information, and learned adaptive adjacency? The present study addresses this question not by reporting only the best MAE, but through a preregistered design with explicit controls and targeted ablation comparisons intended to isolate the contribution of physical road structure.

3. Study Design and Preregistration

Preregistering a study gains real value only if we resist the temptation to abandon its rules when the first practical obstacle appears. I encountered exactly this problem when I attempted to execute the original registered protocol. The seasonal historical data that I had reasonably expected to find in the Istanbul Metropolitan Municipality open data portal were not, in practice, available in the form required by the original plan.

At that point, I had two choices as a researcher. I could quietly substitute different months and proceed as though they had been part of the plan from the beginning, or I could document the problem transparently and modify only the part of the protocol that had become operationally impossible. I chose the second option.

It was particularly important to me that this correction should occur without looking at potential confirmatory outcomes. The availability problem was identified entirely by examining metadata and the resource structure of the official İBB data portal. Traffic speed values from replacement months were neither downloaded nor inspected during this audit. Protocol v2 was therefore not an attempt to modify the hypothesis after seeing an inconvenient result. It was a feasibility correction intended to make execution of the registered study possible while preserving its original scientific question.

During the transition from the original protocol to Protocol v2, the scientific structure of the experiment remained intact. The four hypotheses, H1 through H4, the model conditions, the 64-node design, input history, forecast horizons, fixed hyperparameters, random seeds, chronological data splitting, paired Wilcoxon tests, hierarchical bootstrap procedure, and Holm multiplicity correction remained unchanged in principle. No model, seed, hyperparameter, graph condition, statistical test, or hypothesis was changed in response to confirmatory performance.

Protocol v2 defined seasonal candidate and fallback months using resources visible in the official metadata. After the preregistered availability and coverage criteria were applied, May and November 2024 were the two months that qualified for confirmatory inference. February and August 2024 failed the requirement that at least 64 locations achieve 98% training-period unique-hour coverage.

I also made a deliberate decision to exclude January 2025 from confirmatory inference, even though data from that month were available. The reason was methodological rather than seasonal. January 2025 had already been used during exploratory development for pipeline validation, sampling experiments, graph ablations, road-graph comparisons, and residual graph-signal analyses. Reusing it for confirmatory testing would have weakened the separation between exploration and confirmation. I therefore kept it outside the confirmatory evidence altogether.

Fixing the rules in advance inevitably reduces a researcher's freedom during execution. It prevents us from moving easily toward a more attractive result when the evidence does not behave as expected. But that restriction is precisely what gives the final result its value. When I eventually examined the confirmatory numbers, I wanted to know that I was looking at what the data actually said, rather than at a result gradually shaped by my own engineering expectations.

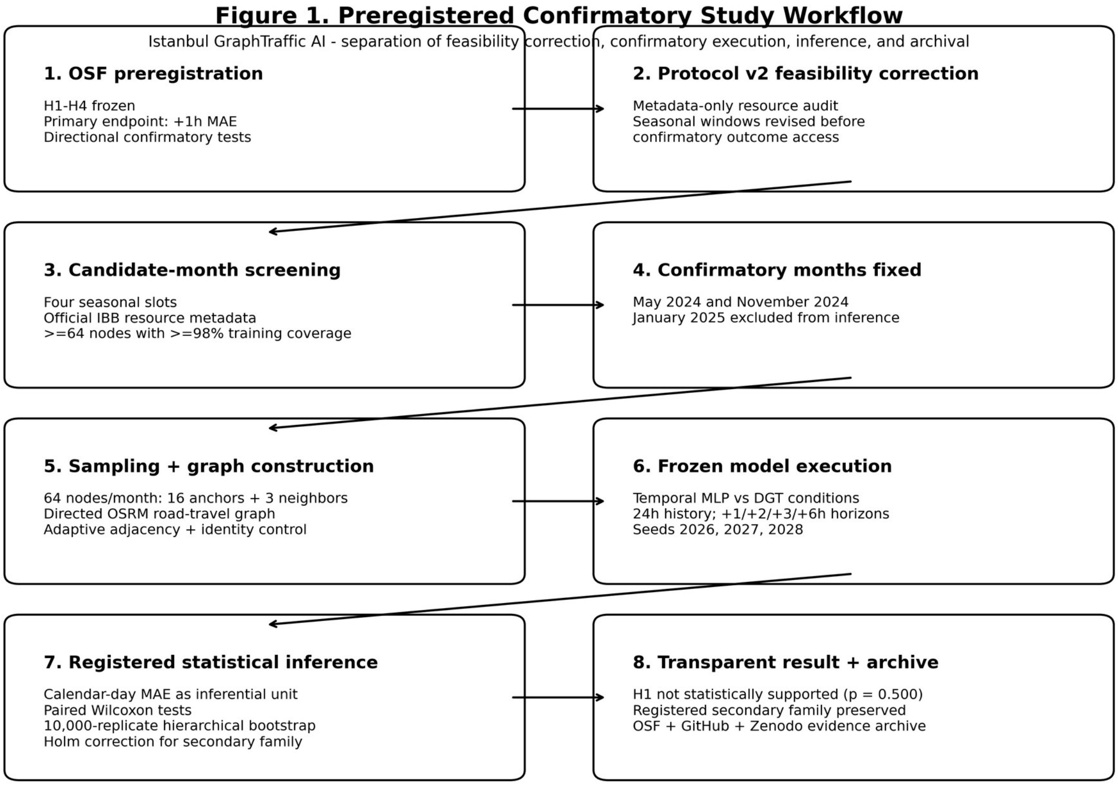

Figure 1. Preregistered confirmatory study workflow. The figure separates the original OSF preregistration, the metadata-only feasibility correction that produced Protocol v2, seasonal eligibility screening, deterministic 64-node sampling, directed road-graph construction, frozen model execution, registered H1-H4 inference, and public archival. The feasibility correction preceded confirmatory outcome access and did not modify the registered hypotheses, statistical tests, seeds, or model conditions.

4. Data and Node Selection

4.1 Data source and model inputs

The study used the publicly available Istanbul Metropolitan Municipality (İBB) Hourly Traffic Density Data Set. The forecasting target was average traffic speed in the source unit of km/h. Spatial information included traffic-location identifiers and geographic coordinates. The confirmatory input features were standardized speed (speed_z), sine and cosine encodings of hour of day, sine and cosine encodings of day of week, and a weekend indicator. The input-history window was 24 hours, and the registered forecast horizons were +1, +2, +3, and +6 hours.

4.2 Seasonal candidate design

Protocol v2 fixed four seasonal slots and a deterministic candidate and fallback order. Availability was determined from official resource metadata before traffic values from replacement months were downloaded or inspected. A month could be excluded from confirmatory inference only if the official resource was unavailable or fewer than 64 nodes satisfied the preregistered training-only coverage rule.

Seasonal slot	Primary candidate	Fallback
Winter	2024-02	2024-03
Spring	2024-05	2024-04
Summer	2024-08	2024-07
Autumn	2024-11	2024-10

The final analyzable confirmatory months were May 2024 and November 2024. February and August 2024 did not satisfy the preregistered requirement of at least 64 locations with 98% training-period unique-hour coverage. January 2025 was excluded because it had already been used in exploratory development.

4.3 Node eligibility and deterministic selection

Once I had worked through the challenge of selecting feasible confirmatory months, the next question was which traffic locations should actually enter the experiment. Istanbul contains a large number of traffic measurement points spread across a vast urban area, and selecting only a subset of them creates another opportunity for researcher choice.

This was exactly the kind of choice I wanted to constrain. It would have been easy to favor locations with cleaner traffic patterns, more convenient behavior, or characteristics that appeared more suitable for graph-based modeling. Instead, I used a deliberately outcome-blind eligibility rule. For each confirmatory month, node eligibility was determined using only the training partition. Information from the validation and test periods played no role in deciding which locations were allowed into the experiment.

The first requirement was data coverage. A location had to contain at least 98% unique-hour coverage during the training period. Locations with insufficient temporal coverage were excluded regardless of what their traffic behavior might otherwise have looked like.

After this initial screening, I faced a second problem. I did not want the selected network to become concentrated around one busy neighborhood, one highway corridor, or one geographically narrow part of Istanbul. At the same time, a graph-based experiment needs enough local density for meaningful spatial relationships to exist. I therefore used a citywide-anchor and local-neighborhood design.

For each confirmatory month, sixteen geographically diverse eligible locations were selected as anchors. Around each anchor, the three nearest unused eligible locations were then added. This produced sixteen local groups of four locations and a total of 64 nodes for that month. The design was intended to balance two competing needs: broad geographic diversity across Istanbul and sufficient local concentration for graph-based spatial modeling.

One rule in this selection process was especially important to me. Traffic speed and vehicle volume were not used to decide which locations entered the 64-node network. I did not choose a location because it represented a persistently congested intersection, an unusually smooth road segment, or a traffic pattern that appeared favorable to one model. Those outcome-related characteristics were deliberately excluded from the selection decision.

For me, this was an important safeguard against creating a playing field that was already tilted toward a particular model. The resulting 64-node design combined citywide coverage with local neighborhoods while keeping the actual traffic outcomes outside the node-selection process.

5. Graph Construction

5.1 Directed road-travel graph

We now arrive at one of the most personally compelling parts of this study. Having spent years moving between the logic of artificial intelligence engineering and the world of language, I found that converting Istanbul's busy streets into a numerical graph felt almost like translating a complex urban story into the language of a machine.

A common shortcut in spatial modeling is to assume that two locations that are geographically close on a map should also be strongly related. In a city such as Istanbul, straight-line distance can be misleading. One-way streets, road configuration, direction of travel, and the actual routes available to vehicles can make two geographically close points very different from two points that are genuinely close in terms of travel. I therefore wanted the graph to reflect the physical road network and real driving relationships rather than simple geometric proximity.

For this reason, I used routing information derived from OpenStreetMap and OSRM. For each of the 64 nodes, I asked a practical question: if a driver actually departed from this location, which other selected locations would be closest in terms of travel time? Geographic distance alone was not the criterion. Among all possible destinations, I retained the six outgoing neighbors with the lowest travel times.

This is also what made the graph directed. In a real urban network, traveling from A to B is not necessarily equivalent to traveling from B to A. Road direction, routing constraints, and network configuration can produce different travel times in opposite directions. The relationships represented in the graph were therefore allowed to be asymmetric and directional.

The strength of each relationship was continuous rather than simply present or absent. Shorter travel times represented stronger connections. Directed edge weights were defined using an exponential decay function:

$$w\_ij = exp(-t\_ij / \tau)$$

Here, t_ij is the directed travel time in seconds from node i to node j, while τ was defined as the median selected outgoing travel time for the corresponding month. In this way, the weighting scale was derived from the travel-time structure of that month's selected network.

5.2 Adaptive adjacency

Even with this carefully constructed physical graph, I did not want to force the model to rely only on relationships that I had predefined. A physical road graph describes the transportation network, but traffic dependencies can also arise from patterns that are not obvious from physical connectivity alone. I therefore allowed the DGT to learn a second graph directly from the data through adaptive adjacency. Trainable source and destination node embeddings produced a learned affinity matrix through a sigmoid transformation of scaled pairwise embedding products.

The model consequently had access both to the explicitly constructed physical road network and to relationships that it could learn for itself. The normalized static adjacency and learned adaptive adjacency were combined using a fixed mixture:

$$A\_combined = clip(0.65\ A\_static + 0.35\ A\_adaptive,\ 0,\ 1)$$

The combined matrix was transformed with a log operation into an additive attention mask before multi-head attention. The 0.65 static and 0.35 adaptive mixture was frozen in Protocol v2.

5.3 Identity control graph

Finally, I needed a strict control condition to determine whether the effort invested in constructing the physical road graph actually contributed anything measurable. I therefore kept the DGT architecture and its adaptive adjacency mechanism unchanged, but replaced the physical road graph with an identity matrix. In that control condition, the static graph provided no cross-node physical connectivity. This comparison allowed me to ask the question at the heart of the study: when the physical road graph is removed while the rest of the graph model remains intact, does predictive performance meaningfully change?

6. Forecasting Models

6.1 Temporal MLP

In engineering and science, the strength of a model is best judged against a serious opponent. If the goal is to understand the added value of spatial information and the road graph, the graph model should not be tested against a weak baseline. If a Dynamic Graph Transformer defeats a poorly equipped competitor, that result does not demonstrate the value of the graph. It only demonstrates the weakness of the opponent. For this reason, I needed a strong temporal comparator.

That competitor was the Temporal MLP. It was deliberately blind to location and received no spatial adjacency information, but it was given a strong temporal view of the problem. For each node, I supplied the previous 24 hours of traffic history together with the temporal input features. These inputs were flattened and passed through a hidden layer with dimension 48. ReLU activation and a dropout rate of 0.10 were used to provide nonlinear capacity while limiting overfitting. With this structure, the model directly predicted the required forecast horizons.

6.2 Dynamic Graph Transformer

On the other side stood the more complex DGT, designed to understand both time and space. The model began with an input projection that transformed the features of each node. A GRU then summarized the information contained in the previous 24 hours. Once this temporal history had been encoded, the node representations entered a four-head graph-masked self-attention mechanism. Residual connections, LayerNorm, and a feed-forward block followed the attention layer.

The output of the DGT was not limited to a single point prediction. It produced predictive quantiles at 0.1, 0.5, and 0.9 for each horizon and node. To prevent quantile crossing and preserve their logical ordering, the output quantiles were sorted before the final forecasts were used.

The central logic of the ablation can therefore be summarized through three model conditions. The Temporal MLP represents temporal information without spatial adjacency. DGT road + adaptive represents temporal information, explicit physical spatial structure, and learned relationships. DGT identity + adaptive retains temporal modeling, graph architecture, and learned adaptive relationships, but removes the actual physical road graph. The comparison among these three conditions is what allows the study to ask whether explicit road structure contributes additional predictive information.

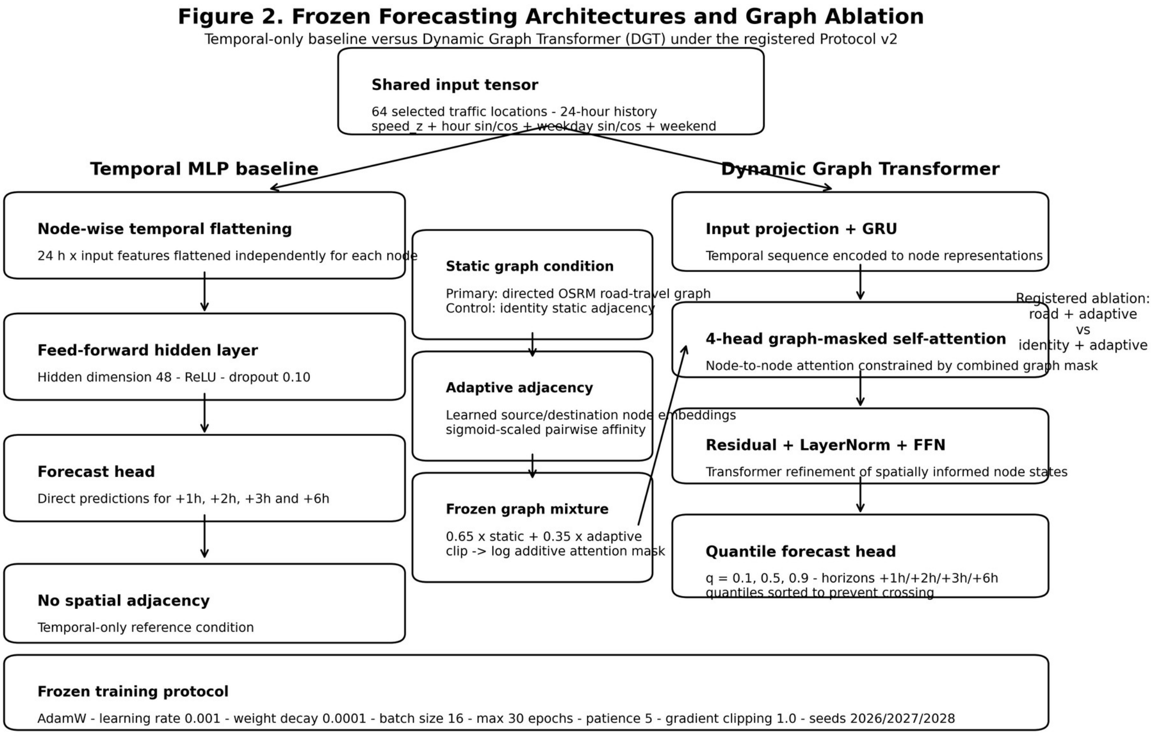

Figure 2. Frozen forecasting architectures and graph ablation. The Temporal MLP uses node-wise temporal history without spatial adjacency. The Dynamic Graph Transformer combines GRU temporal encoding, four-head graph-masked self-attention, residual and normalization processing, and quantile forecasting. The registered physical-graph ablation compares directed road-travel static adjacency plus adaptive adjacency against identity static adjacency plus the same adaptive mechanism while keeping the rest of the DGT condition fixed.

6.3 Historical-average descriptive baseline

Protocol v2 also included a deterministic historical-average baseline. It predicts each node using the training-period average traffic speed for the corresponding hour of week, with the global training mean as fallback. This baseline was retained for descriptive benchmarking and reproducibility, but it was not part of the preregistered H1-H4 inferential comparisons.

6.4 Frozen hyperparameters

Hyperparameter	Value
Hidden dimension	48
Attention heads	4
Dropout	0.10
Maximum epochs	30
Batch size	16
Learning rate	0.001
Weight decay	0.0001
Early-stopping patience	5
Random seeds	2026, 2027, 2028

Models were trained with AdamW. DGT quantile outputs used pinball loss, while non-quantile outputs used L1 loss. Gradient norms were clipped at 1.0. Model selection used validation loss with early stopping.

7. Experimental Protocol and Statistical Analysis

7.1 Chronological splitting and fixed seeds

When two models enter the ring, architecture is not the only thing that matters. Equally important is the integrity of the experimental protocol used to judge them. In time-series forecasting, time moves in only one direction. We cannot place an entire month of traffic observations into a bag, shuffle them randomly, and feed the resulting mixture to the model. Allowing future information to leak into the past would create a deceptively impressive but invalid experiment.

To preserve temporal integrity, I divided each confirmatory month chronologically into three consecutive partitions: 70% for training, 15% for validation, and the final 15% for testing. No random temporal shuffling was used. The past remained in the past, and the future remained unseen until evaluation.

Neural networks can also appear unusually strong because of a fortunate initialization. To reduce the influence of such chance, I trained each neural model under three fixed and predetermined random seeds: 2026, 2027, and 2028. I did not want a conclusion to depend on a single favorable initialization.

7.2 Statistical unit of analysis

A particularly important part of the protocol concerned the statistical unit of analysis. I did not treat every individual prediction or every traffic node as an independent inferential observation. For each test calendar day, MAE was first aggregated across the 64 nodes and then averaged across the three neural seeds. The resulting calendar-day value became the unit used for paired statistical inference. Across the two analyzable confirmatory months, this produced 10 paired calendar-day units. The resulting statistical sample was small, but it avoided artificially inflating the number of independent observations.

7.3 Registered hypotheses and tests

The primary hypothesis, H1, defined the main contest of the study. At the +1 hour horizon, I asked whether the DGT equipped with directed road adjacency and adaptive adjacency achieved lower held-out MAE than the Temporal MLP. Because the directional alternative had been specified in advance, H1 was evaluated using a one-sided paired Wilcoxon signed-rank test with $\alpha = 0.05$. H1 was the single primary confirmatory test and therefore was not subject to multiplicity adjustment.

The remaining registered hypotheses were treated differently. H2, H3, and the three horizon-specific H4 comparisons formed one secondary inferential family. Their p values were adjusted using Holm's procedure so that repeated testing within this family would not create an artificially permissive standard for statistical significance.

- H1, primary and directional: at +1h, DGT directed-road + adaptive has lower held-out MAE than Temporal MLP.
- H2, secondary and directional: at +1h, DGT directed-road + adaptive has lower held-out MAE than DGT identity + adaptive.
- H3, secondary and directional: the incremental advantage of DGT directed-road + adaptive over Temporal MLP is larger at +1h than at +6h.
- H4, secondary and directional: at +2h, +3h, and +6h, DGT directed-road + adaptive has lower held-out MAE than Temporal MLP.

7.4 Effect sizes and uncertainty

The primary effect measure was the mean paired daily MAE difference, Δ = MAE_model - MAE_reference. Negative values therefore favor the registered graph-model effect. Relative MAE difference was also reported.

Uncertainty was quantified using a hierarchical bootstrap with 10,000 replicates and a fixed bootstrap seed of 20260814. The resampling respected the hierarchy of the confirmatory design by resampling months first and then test days within sampled months. The resulting 95% bootstrap intervals were used for effect estimation. These intervals were not multiplicity-adjusted. For the secondary hypothesis family, preregistered statistical decisions were based on Holm-adjusted p values rather than on whether an individual unadjusted bootstrap interval happened to exclude zero.

7.5 Registered diagnostic endpoints and reporting boundary

Protocol v2 additionally listed RMSE, MAPE, R^2, predictive-interval coverage, predictive-interval width, and random or structured sensor-failure robustness as secondary diagnostics. The archived confirmatory metrics contain RMSE, MAPE, R^2, and, for quantile models, predictive-interval coverage and width. These diagnostics do not replace, redefine, or overturn H1-H4. Sensor-failure robustness results are not part of the archived confirmatory evidence package used for this manuscript and are therefore not claimed here.

8. Results

8.1 Confirmatory feasibility

Two seasonal months satisfied the preregistered resource and node-coverage criteria: May 2024 and November 2024. The minimum requirement of at least two analyzable seasonal slots was therefore met, and confirmatory inferential testing proceeded.

8.2 Primary H1 result

After all the care invested in the experimental design, preregistration, and locking the rules of the study in advance, the moment finally arrived when I was allowed to look at the results. From an engineering perspective, I had expected the physical road graph to leave a clear footprint, at least at the one-hour forecasting horizon. But numbers have no obligation to satisfy our expectations.

I began with H1, the main contest of the study. The DGT equipped with directed road adjacency and adaptive adjacency achieved an MAE of 3.6804 km/h. The Temporal MLP achieved 3.6986 km/h. The paired difference was -0.0181 km/h, corresponding to an estimated improvement of approximately 0.490% in favor of the graph model. This numerical advantage was not statistically reliable. The 95% hierarchical bootstrap confidence interval extended from -0.1760 to +0.1120 km/h and included zero, while the preregistered one-sided p value was 0.5000. H1 was therefore not statistically supported.

8.3 Secondary H2 and H4 results

The result for H2 was, in some ways, even more revealing. When the DGT containing the physical road graph was compared with the control DGT using identity static adjacency, the road-graph condition was not better. Its point estimate was slightly worse. The paired difference was +0.0207 km/h, corresponding to +0.565% relative to the control condition. The raw p value was 0.903320 and the Holm-adjusted p value was 1.000000. H2 therefore provided no statistical support for an incremental benefit from the directed physical road graph.

The additional H4 forecast horizons also showed no consistent statistically supported advantage. At +2h, the paired difference was +0.0327 km/h, favoring the Temporal MLP. At +3h, the difference was -0.0390 km/h, favoring the DGT. At +6h, the difference was +0.1392 km/h, again favoring the Temporal MLP. None of these horizon-specific differences was statistically supported under the registered analysis.

8.4 H3 short-horizon versus long-horizon contrast

The most intriguing and potentially misleading result appeared in H3, where I asked whether the graph model's relative advantage was greater at +1h than at +6h. The estimated contrast was -0.1574 km/h. Its 95% hierarchical bootstrap interval ranged from -0.2539 to -0.0383 km/h, and the raw p value was 0.013672. At first glance, this is exactly the kind of result that tempts a researcher to declare statistical significance. But H3 belonged to the registered secondary hypothesis family, so the multiplicity correction could not be ignored. After applying the preregistered Holm procedure, the adjusted p value became 0.068359. H3 therefore did not meet the registered familywise significance criterion.

Numbers sometimes offer an attractive signal from inside a table, but the researcher's job is not to treat every promising signal as proof. Under the rules fixed before examining the confirmatory outcomes, none of the registered superiority hypotheses obtained sufficient statistical evidence for confirmation.

Hypothesis	Comparison	Horizon	Mean paired MAE difference (km/h)	Relative difference	95% bootstrap CI (km/h)	Raw p	Holm-adjusted p
H1	DGT road+adaptive vs Temporal MLP	+1h	-0.0181	-0.490%	[-0.1760, +0.1120]	0.500000	-

Hypothesis	**Comparison**	**Horizon**	**Mean paired MAE difference (km/h)**	**Relative difference**	**95% bootstrap CI (km/h)**	**Raw p**	**Holm-adjusted p**
H2	DGT road+adaptive vs DGT identity+adaptive	+1h	+0.0207	+0.565%	[-0.0281, +0.0700]	0.903320	1.000000
H4_2h	DGT road+adaptive vs Temporal MLP	+2h	+0.0327	+0.852%	[-0.1085, +0.1575]	0.753906	1.000000
H4_3h	DGT road+adaptive vs Temporal MLP	+3h	-0.0390	-0.979%	[-0.1495, +0.0376]	0.347656	1.000000
H4_6h	DGT road+adaptive vs Temporal MLP	+6h	+0.1392	+3.391%	[-0.1086, +0.3518]	0.975586	1.000000
H3	(+1h effect) vs (+6h effect)	+1h vs +6h	-0.1574	-3.881%	[-0.2539, -0.0383]	0.013672	0.068359

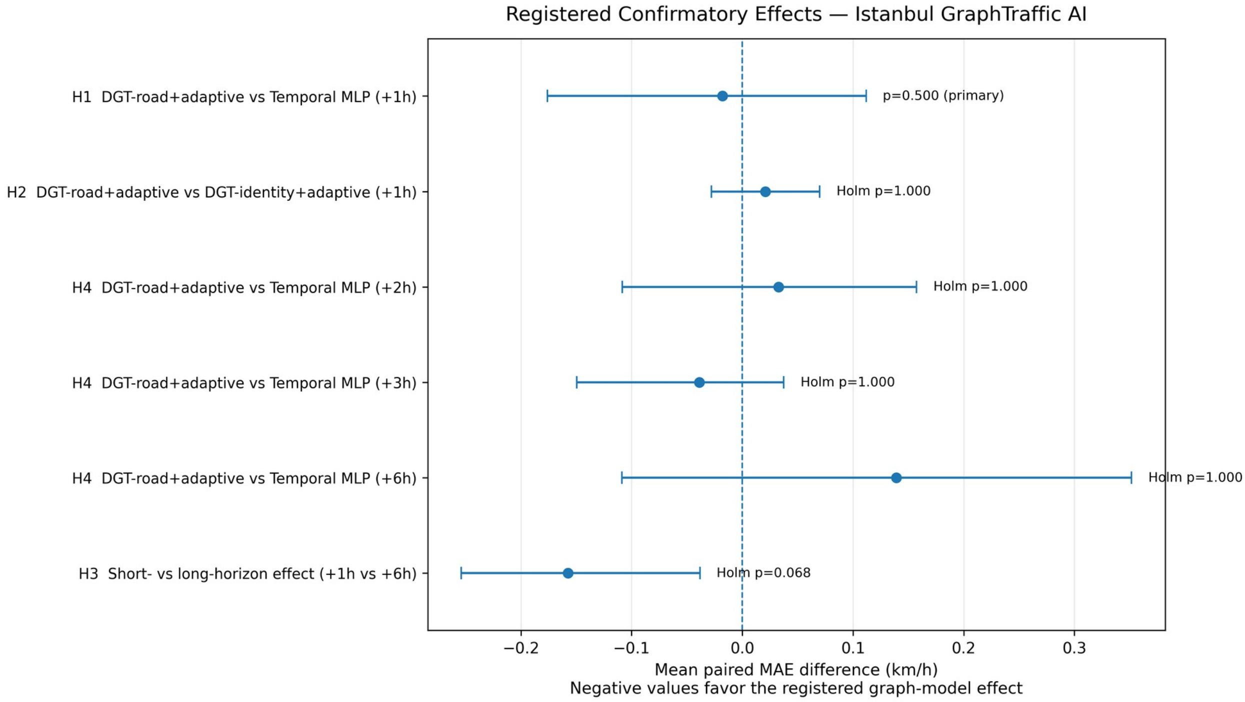

Figure 3. Registered confirmatory effects. Mean paired MAE differences are shown with 95% hierarchical-bootstrap confidence intervals. Negative differences favor the registered graph-model effect. H1 was the unadjusted primary test; H2, H3, and the three H4 horizon-specific tests belonged to the Holm-adjusted secondary family. Bootstrap intervals are unadjusted effect-estimation intervals. Statistical decisions for the secondary family use Holm-adjusted p values, so H3 does not meet the preregistered familywise significance criterion despite its unadjusted interval excluding zero.

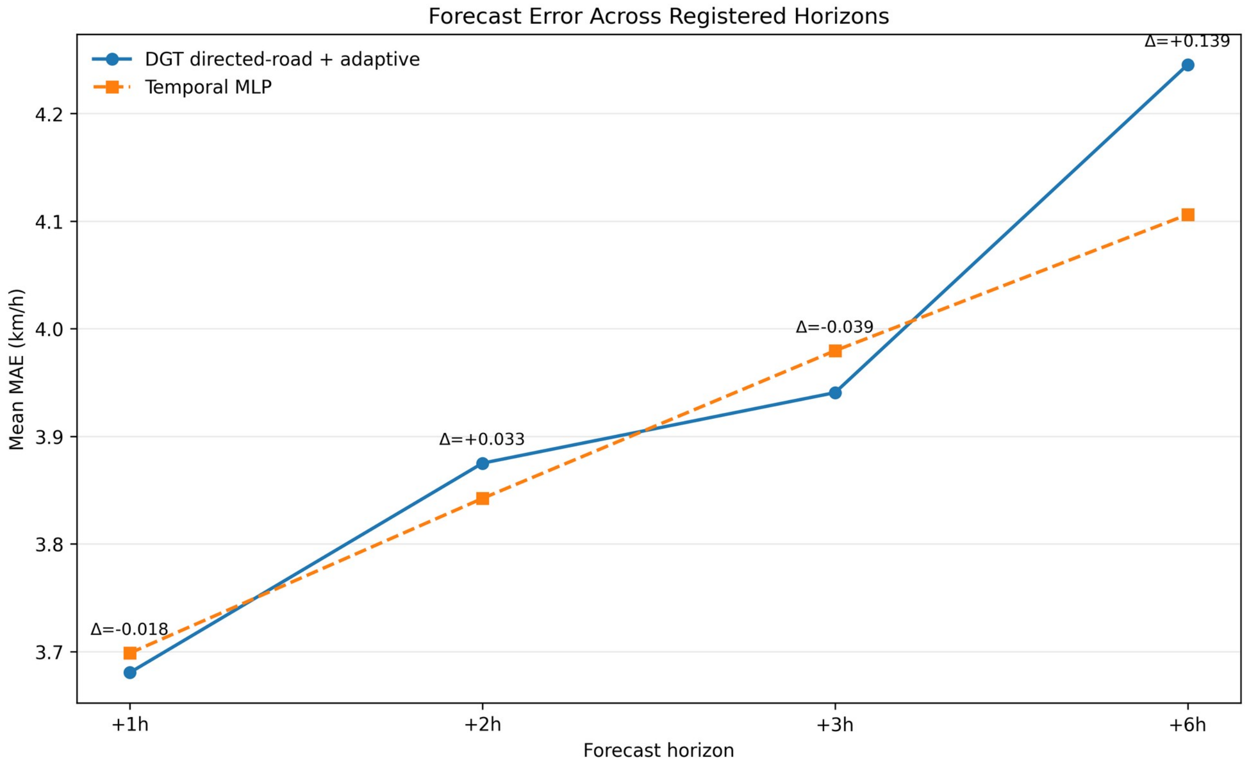

Figure 4. Mean forecast error across registered horizons. Mean MAE for DGT directed-road + adaptive and Temporal MLP is shown at +1h, +2h, +3h, and +6h. The annotated differences are DGT minus Temporal MLP. Negative values favor DGT. The graph model was numerically better at +1h and +3h and worse at +2h and +6h, but none of the horizon-specific registered superiority comparisons was statistically supported.

8.5 Interpretation boundary

The confirmatory evidence does not support a claim that the directed-road DGT is superior to the Temporal MLP under the tested design. It also does not establish that road graphs are generally unhelpful for traffic forecasting. The inference is narrower: under these two analyzable Istanbul months, the frozen 64-node sampling design, fixed architecture, fixed hyperparameters, and preregistered statistical procedure, directed road-travel structure did not yield a reliable incremental improvement in MAE.

9. Discussion

Perhaps the most interesting and meaningful finding of this study is precisely what we did not find. I entered the experiment with a strong engineering intuition: if I gave the model the real physical road network and allowed it to see the structure of the city, surely it should become a better and more informed forecaster. In the end, however, the numbers forced me to draw a clear line between the elegance of an idea on paper and the evidence required to support it.

Why, then, did the carefully constructed road graph fail to produce the substantial advantage I had expected? Several explanations seem plausible to me. First, the 24-hour temporal history may already have provided such a rich and powerful predictive signal that the model needed relatively little additional help from explicit spatial structure. Second, the DGT was also allowed to learn through adaptive adjacency. This learned component may have recovered a substantial part of the hidden spatial relationships directly from the data without requiring those relationships to be fully specified by the physical road map. Third, a network of 64 nodes represents only a limited view of a city as large and interconnected as Istanbul. Important traffic propagation paths may pass through locations that were not represented in the selected network, making the observed graph too sparse to capture the full spatial chain of urban traffic. Finally, the forecasting target was based on hourly traffic measurements. Rapid propagation effects that develop over several minutes may be weakened or lost when traffic behavior is represented at an hourly resolution.

What, then, should we make of H3? There was a suggestive pattern indicating that spatial information might have been relatively more useful at the shorter, one-hour horizon than at the six-hour horizon. But once the preregistered Holm correction was applied, the adjusted p value was 0.0684. A promising pattern is not the same as a confirmed discovery. I see H3 as a footprint in the snow: a potentially meaningful pattern that may guide future replication and investigation, but not a flag that can yet be planted as evidence of a confirmed effect.

Were these null results a failure for me as a researcher? I do not see them that way. This is precisely where preregistration demonstrated its value. Without the constraints established in advance, it would have been easy to change the month, remove inconvenient nodes, try different random seeds, alter the architecture, or emphasize another forecast horizon until a statistically attractive result appeared. The value of this experiment lies in resisting that temptation. When the confirmatory evidence gave my original expectation a clear no, I reported that no rather than trying to transform it into a more appealing answer. To me, that is an important part of research maturity.

The boundary of my claim is deliberately narrow. This study does not show that graph neural networks are useless for traffic forecasting. It does not establish a general ranking of graph and non-graph approaches. The conclusion is limited to the experimental conditions that were actually tested. Across these two confirmatory months, with the selected 64-node design, the specified DGT architecture, the frozen hyperparameters, and the preregistered statistical protocol, adding explicit directed road structure did not produce a reliable incremental reduction in MAE. The broader question of when physical road graphs provide meaningful predictive value remains open for future models, larger networks, different temporal resolutions, and independent replications.

10. Limitations

Every scientific experiment is like a small window looking out onto a much larger landscape. It is never the entire landscape. If I fail to observe something through the particular window used in this study, I cannot conclude that it does not exist elsewhere in the city or under different conditions. Clearly stating the limitations of a study is part of what gives research credibility. The boundaries of the present findings should therefore be made explicit.

10.1 Only two seasonal periods were analyzable

I would have preferred to have evidence representing all four seasonal slots of Istanbul. Protocol v2 was designed around four seasonal candidate periods. In the final execution, however, only May and November 2024 satisfied the preregistered availability and coverage requirements. The seasonal evidence is therefore narrower than originally intended, and the results should not be treated as representing the full range of seasonal traffic behavior in Istanbul.

10.2 Limited statistical power

After applying the registered aggregation procedure, the inferential analysis contained only 10 paired calendar-day units. This restricts the statistical power available for detecting small effects. The null findings in this study therefore should not be interpreted as evidence that physical road structure has absolutely no predictive effect. The more precise conclusion is that, with the available confirmatory sample, I did not obtain sufficient evidence for a reliable incremental effect.

10.3 A 64-node view of a much larger city

Istanbul contains a far larger and more interconnected transportation network than the 64 locations included in each monthly experiment. The physical road graph may carry useful information that was only partially observable through this selected network. Traffic propagation between two observed locations may pass through intermediate parts of the city that were not represented among the selected nodes. The 64-node design should therefore be understood as a controlled sample of the urban network rather than a complete representation of Istanbul traffic.

10.4 Month-specific graph structure

The directed road graph and its travel-time scale parameter, τ, were constructed separately for the selected nodes of each confirmatory month. As a result, the detailed topology presented to the model in May was not necessarily identical to that used in November. This is not inherently a methodological defect, but it limits direct fine-grained comparison of graph structure across the two months.

10.5 Hourly temporal resolution

The forecasting target was hourly average traffic speed. Traffic dynamics can evolve on much shorter time scales. If the strongest contribution of physical road connectivity occurs through propagation over intervals of five, ten, or fifteen minutes, hourly aggregation may weaken or obscure those effects. The present result therefore should not automatically be generalized to traffic data at finer temporal resolutions.

10.6 The study does not judge all graph architectures

This experiment evaluated a specific Temporal MLP and a specific DGT implementation under frozen hyperparameters. It does not establish a general ranking of graph neural networks against temporal models. Different architectures, larger node sets, finer temporal resolution, alternative graph-construction methods, or different training regimes could produce different results.

10.7 This was not a state-of-the-art benchmark competition

The purpose of the study was not to achieve the lowest possible MAE by comparing every recent traffic-forecasting architecture. The research question was narrower: how much incremental predictive value does explicit physical road structure contribute under a controlled comparison? The findings should therefore be interpreted as evidence from a targeted ablation study rather than as a claim about overall state-of-the-art forecasting performance.

10.8 Prior exploratory development

January 2025 data were used during earlier exploratory development of the project, and that exploratory work contributed to shaping and refining the eventual research question and design. For this reason, January 2025 was deliberately excluded from confirmatory testing. The replacement confirmatory months were not inspected for traffic outcomes before Protocol v2 was fixed. Transparency about prior exploration strengthens the interpretation of the later confirmatory analysis.

11. Conclusion

In the end, all of the analysis, ablation, and argument with the numbers brings us back to the same simple question that started this study. In the noisy and deeply interconnected world of urban traffic, when we already have a strong temporal model that understands recurring time patterns and an adaptive adjacency mechanism capable of learning hidden relationships from the data, does explicitly adding the physical road network still provide reliable incremental predictive value?

Under the frozen conditions of this experiment, the answer was that we did not find sufficient evidence for such an advantage. The primary hypothesis, H1, produced only a small difference of -0.0181 km/h, corresponding to approximately 0.49% in favor of the DGT. With a p value of 0.5000, that difference was far too uncertain to support a claim of reliable superiority. More broadly, none of the preregistered superiority hypotheses obtained sufficient statistical evidence for confirmation.

This null result should not be interpreted in absolute terms. It does not mean that physical road graphs are useless, that graph neural networks have followed the wrong direction, or that temporal models will always outperform graph-based approaches. The conclusion is much narrower. Under the specific data, architecture, graph construction, and preregistered protocol used in this study, I could not demonstrate a reliable incremental advantage from explicitly providing the directed physical road structure.

Perhaps the larger lesson of this work extends beyond traffic forecasting and the streets of Istanbul. As engineers, we are often attracted to complexity. We tend to assume that a structure that is more sophisticated, more elegant, or more impressive on paper must also carry more useful information. But complexity earns its value only when its independent contribution can be measured against a serious baseline. For me, one of the most important acts in building an intelligent model is not simply deciding what else to add. It is having the courage to remove a beautiful and complicated component and ask a much harder question: if I take this part away, what do I actually lose in my understanding of the problem?

12. Reproducibility and Research Integrity

The project preserves a public workflow intended to make the confirmatory analysis independently auditable. The archived evidence includes:

- frozen Protocol v2 configuration;
- OSF preregistration and public update;
- resource-availability audit;

- selected-node files;
- graph diagnostics;
- OSRM routing matrices and responses;
- raw-source provenance;
- fixed random seeds;
- model code;
- confirmatory metrics;
- daily paired errors;
- registered effect table;
- bootstrap metadata;
- hypothesis-test results;
- SHA-256 artifact checksums.

The registered confirmatory result is archived independently of whether it supports the original hypothesis. Synthetic data in the repository are used only for pipeline smoke testing and are not treated as evidence for the real-data confirmatory conclusions.

13. Data and Code Availability

Public project repository: https://github.com/FaramarzKowsari/istanbul-graphtraffic-ai

Research website: https://faramarzkowsari.github.io/istanbul-graphtraffic-ai/

Registered confirmatory results: https://faramarzkowsari.github.io/istanbul-graphtraffic-ai/confirmatory-results.html

Machine-readable confirmatory dataset landing page: https://faramarzkowsari.github.io/istanbul-graphtraffic-ai/registered-confirmatory-dataset.html

OSF registration: https://doi.org/10.17605/OSF.IO/FM5R7

Versioned project archive: https://doi.org/10.5281/zenodo.21916357

The underlying İBB source data remain subject to the source provider's terms and availability. Derived confirmatory outputs published by this project are distributed under CC BY 4.0 where indicated.

14. Ethics Statement

The study analyzes publicly available aggregate traffic data and does not involve human participants, clinical data, personal identifiers, or intervention on individuals. No human-subject experimentation was conducted.

15. Funding

No external funding is declared for this study.

16. Competing Interests

The author declares no competing interests.

17. Author Contributions

Faramarz Kowsari: conceptualization, methodology, software, data curation, formal analysis, validation, visualization, reproducibility engineering, writing - original draft, writing - review and editing, and project administration.

18. Acknowledgments

The author acknowledges Istanbul Metropolitan Municipality for making the traffic data publicly accessible and the open-source communities supporting Python, PyTorch, OpenStreetMap, OSRM, and the broader scientific software ecosystem.

19. AI Assistance Disclosure

AI-assisted tools were used for language translation from author-written Persian drafts into English and for copy-editing and document formatting. The research question, preregistration, study design, software, data processing, statistical analysis, results, interpretations, and substantive scientific claims are the author's. Final responsibility for the manuscript and all claims rests with the author.

References

[1] Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2018). Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. International Conference on Learning Representations (ICLR). arXiv:1707.01926.

[2] Yu, B., Yin, H., & Zhu, Z. (2018). Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 3634-3640. https://doi.org/10.24963/ijcai.2018/505

[3] Guo, S., Lin, Y., Feng, N., Song, C., & Wan, H. (2019). Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 922-929. https://doi.org/10.1609/aaai.v33i01.3301922

[4] Wu, Z., Pan, S., Long, G., Jiang, J., & Zhang, C. (2019). Graph WaveNet for Deep Spatial-Temporal Graph Modeling. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 1907-1913. https://doi.org/10.24963/ijcai.2019/264

[5] Zheng, C., Fan, X., Wang, C., & Qi, J. (2020). GMAN: A Graph Multi-Attention Network for Traffic Prediction. Proceedings of the AAAI Conference on Artificial Intelligence, 34(01), 1234-1241. https://doi.org/10.1609/aaai.v34i01.5477

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 30.

[7] Huang, S., Song, H., Jiang, T., Telikani, A., Shen, J., Zhou, Q., Yong, B., & Wu, Q. (2024). DST-GTN: Dynamic Spatio-Temporal Graph Transformer Network for Traffic Forecasting. arXiv:2404.11996.

[8] Olug, E., Kaya, K., Tugay, R., & Oguducu, S. G. (2024). IBB Traffic Graph Data: Benchmarking and Road Traffic Prediction Model. 2024 IEEE 29th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD). https://doi.org/10.1109/CAMAD62243.2024.10943048

[9] Zhou, J., Liu, E., Chen, W., Zhong, S., & Liang, Y. (2024). Navigating Spatio-Temporal Heterogeneity: A Graph Transformer Approach for Traffic Forecasting. arXiv:2408.10822.

[10] Wang, H., Chen, J., Pan, T., Dong, Z., Zhang, L., Jiang, R., & Song, X. (2024). STGformer: Efficient Spatiotemporal Graph Transformer for Traffic Forecasting. arXiv:2410.00385.

[11] Yang, S., Huang, Z., Wu, Q., & Zhuo, Z. (2025). General Decoupled Graph Convolutional Recurrent Network for Traffic Prediction. IEEE Sensors Journal, 25(18), 35460-35478. https://doi.org/10.1109/JSEN.2025.3580440

[12] Zhao, Z., Shen, G., Zhou, W., Qi, J., Liu, Y., & Kong, X. (2026). Generative Adversarial Pre-Training Enabled Spatial-Temporal Graph Modeling of Traffic Data. IEEE Transactions on Intelligent Transportation Systems, 27(2), 2056-2071. https://doi.org/10.1109/TITS.2025.3633688

Appendix A. Registered Hypotheses

H1 - Primary, directional. At +1h, DGT directed-road + adaptive has lower held-out MAE than Temporal MLP.

H2 - Secondary, directional. At +1h, DGT directed-road + adaptive has lower held-out MAE than DGT identity + adaptive.

H3 - Secondary, directional. The incremental predictive benefit of DGT directed-road + adaptive relative to Temporal MLP is larger at +1h than at +6h.

H4 - Secondary, directional. At +2h, +3h, and +6h, DGT directed-road + adaptive has lower held-out MAE than Temporal MLP.

Appendix B. Confirmatory Effect Table

Hypothesis	Horizon	Mean model MAE	Mean reference MAE	Paired difference	Relative difference	95% CI	Raw p	Holm p
H1	+1h	3.6804	3.6986	-0.0181	-0.490%	[-0.1760, +0.1120]	0.500000	-
H2	+1h	3.6804	3.6597	+0.0207	+0.565%	[-0.0281, +0.0700]	0.903320	1.000000
H4_2h	+2h	3.8750	3.8423	+0.0327	+0.852%	[-0.1085, +0.1575]	0.753906	1.000000
H4_3h	+3h	3.9406	3.9795	-0.0390	-0.979%	[-0.1495, +0.0376]	0.347656	1.000000
H4_6h	+6h	4.2453	4.1061	+0.1392	+3.391%	[-0.1086, +0.3518]	0.975586	1.000000
H3	+1h vs +6h	-	-	-0.1574	-3.881%	[-0.2539, -0.0383]	0.013672	0.068359