South American Publishing

SAP Transport, Mobility & Society. 2025;4:161
doi: 10.62486/tms2025161

REVIEW

Machine learning for public transport demand forecasting: methods, data and deployment challenges

Aprendizaje automático para la predicción de la demanda de transporte público: métodos, datos y retos de implementación

Carlos Ignacio Ruiz Maldonado[1]

[1] Instituto Técnico en Ciencias Computacionales de Reynosa A.C., Reynosa, México

Submitted: 22-04-2025 | Revised: 26-06-2025 | Accepted: 22-08-2025 | Published: 25-08-2025

Corresponding Author: Carlos Ignacio Ruiz Maldonado

ABSTRACT

Machine learning has become the dominant approach to public transport demand forecasting, yet the literature remains scattered across transport and computing communities. This narrative review synthesizes studies published between 2013 and 2025 on the methods, data sources and deployment challenges of machine learning for transit demand prediction. Searches were run in Scopus, OpenAlex and Crossref, and the synthesis was organized around four thematic axes. The evidence shows a progression from statistical time series models to recurrent, convolutional, graph and transformer architectures, with hybrid and ensemble models generally reporting the lowest errors. Smart card and automatic fare collection records remain the principal data source, complemented by weather, calendar and built environment variables and by emerging sources such as trip planners, Wi-Fi and video. Reported performance is difficult to compare because datasets, horizons and metrics differ across studies, and uncertainty quantification is still rare. The main deployment constraints are real time operation, generalization to new stations and networks, interpretability and model maintenance. Closing these gaps will require shared benchmarks, probabilistic outputs and validation with transit agencies. The review identifies a research agenda connecting accuracy, uncertainty and operational usability.

Keywords: machine learning; deep learning; public transport; demand forecasting; smart card data; transit ridership.

RESUMEN

El aprendizaje automático se ha convertido en el enfoque dominante para predecir la demanda de transporte público, aunque la literatura sigue dispersa entre las comunidades del transporte y la computación. Esta revisión narrativa sintetiza estudios publicados entre 2013 y 2025 sobre los métodos, las fuentes de datos y los retos de implementación del aprendizaje automático en la predicción de la demanda de transporte. Las búsquedas se realizaron en Scopus, OpenAlex y Crossref, y la síntesis se organizó en cuatro ejes temáticos. La evidencia muestra una progresión desde los modelos estadísticos de series temporales hacia arquitecturas recurrentes, convolucionales, de grafos y transformadoras, con los modelos híbridos y ensamblados reportando los menores errores. Los registros de tarjetas inteligentes y de validación automática siguen siendo la principal fuente de datos, complementados por variables meteorológicas, de calendario y de entorno construido y por fuentes emergentes como planificadores de viaje, Wi-Fi y vídeo. El rendimiento reportado es difícil de comparar porque los conjuntos de datos, horizontes y métricas difieren entre estudios, y la cuantificación de incertidumbre sigue siendo escasa. Las principales limitaciones para la implementación son la operación en tiempo real, la generalización a estaciones y redes nuevas, la interpretabilidad y el mantenimiento de los modelos. Cerrar estas brechas requerirá referencias comunes, salidas probabilísticas y validación con operadores de transporte.

Palabras clave: aprendizaje automático; aprendizaje profundo; transporte público; predicción de la demanda; datos de tarjetas inteligentes; demanda de pasajeros.

© 2025; The Authors. This Open access article is distributed under CC BY 4.0 licence.

INTRODUCTION

Reliable demand forecasts underpin most decisions of a public transport agency, from timetable design and vehicle dispatching to long term network investment. Errors in these forecasts convert directly into overcrowded vehicles, underused capacity or higher operating costs, and their social consequences fall unevenly on the users who depend on the network most.

The spread of automatic fare collection and automatic vehicle location systems has turned demand forecasting into a data rich problem, and the field has progressively migrated from aggregate statistical models toward learning methods that exploit high resolution, high frequency records. This migration reflects a genuine change in what the data allow.(1,2)

The resulting literature is large but dispersed across two communities that rarely speak to each other. Transport researchers evaluate models against service planning questions, whereas machine learning studies emphasize architectures and benchmark scores, and the two strands use different datasets, metrics and evaluation habits.(3)

This dispersion makes it difficult for an agency or a researcher entering the field to judge which methods are mature, which data are necessary, and which obstacles appear only when a model leaves the laboratory. A synthesis that reads both strands together is therefore timely, and the absence of a widely shared benchmark makes such a synthesis more useful than a mere catalogue of models.

This narrative review synthesizes the evidence published between 2013 and 2025 on machine learning for public transport demand forecasting. It is organized around four thematic axes: the evolution of model families, the data sources and features that feed them, the reported predictive performance and benchmarking practice, and the deployment challenges of real time operation, generalization, interpretability and maintenance. The objective is to map what is established, where results diverge, and where the field should go next.

DEVELOPMENT

The synthesis that follows is organized in four axes. The study counts that close each axis refer to the distinct core studies coded to that axis in the synthesis matrix.

Model families and their evolution

The strongest convergence in the reviewed evidence is the displacement of single equation statistical baselines by deep architectures once high frequency data become available.

Recurrent networks of the long short term memory family and their gated variants dominate short term passenger flow prediction for rail systems, because they retain temporal context across irregular intervals. Studies that compare these networks against autoregressive baselines report lower errors at the station and network level, and convolutional layers are typically added to extract local temporal patterns before the recurrent stage.(8-11)

The lineage of these architectures runs through the wider traffic prediction literature rather than being original to transit. Recurrent and convolutional designs were first validated on road traffic flow and speed, and the transit studies reviewed here adapted them to passenger counts, which are sparser and more tightly bound to schedules than vehicle flows.

This borrowing explains both the speed of progress and a subtle mismatch, because assumptions that hold for dense road sensors do not always transfer to stops that record demand only when a vehicle calls. The axis therefore reads many methodological choices as inherited priors and flags the studies that test those priors rather than import them unchecked.(8,9)

Attention and transformer architectures represent a further step. By weighting all past observations rather than compressing them into a single hidden state, attention mechanisms retain the long range dependencies that matter for event driven demand, and the reviewed transformer based models target precisely the large scale events and multi step horizons where sequential models lose context.(12,13)

The reported gains concentrate in peak periods and in stations with strongly irregular profiles, whereas off peak and stable demand is served almost as well by simpler models.

Graph based models represent the second major advance, and their rationale is spatial. By encoding stations or zones as nodes and their connectivity as edges, graph convolutional and graph attention networks capture dependencies between stops that recurrent and convolutional models treat only indirectly.(14-16)Dynamic and multiview graphs extend this idea to networks whose structure changes over time or whose demand responds to several overlapping relations-

hips at once.(17-19)

A third pattern is the proliferation of hybrid and ensemble designs that stack decomposition, attention and multiple base learners in a single pipeline. Decomposition into trend and seasonal components, followed by separate learners, is the most common recipe, and ensemble variants that combine several base models report further gains.(20-24)

The comparison between pure and hybrid architectures is further complicated by the heterogeneity of evaluation settings, because studies report relative error reductions against statistical baselines, absolute errors on held out periods, or both, and few add computational cost.

The shift toward deep models also changed the division of labour between feature engineering and learning. Early applications hand crafted calendar, weather and network features and fed them to a shallow learner, whereas the deeper architectures absorb part of that engineering into the network itself, in the form of attention weights, graph edges and learned embeddings.(9,22)

Feature engineering did not disappear; it migrated from the input layer to the architecture, and the axis shows no study in which a deep model succeeded without any attention to the temporal and spatial structure of the problem.

Irrespective of these gains, simple learners do not disappear from the evidence. They remain attractive because they train quickly, expose their assumptions and work with short histories, and several of the reviewed studies retain them as a baseline against which a deeper architecture must justify itself.(9,10)

The persistence of these baselines matters for practice, because an agency with limited data science capacity may obtain most of the available benefit from a carefully tuned ensemble rather than from a deep network.

The choice among families also tracks the structure of the problem rather than any universal ranking. Recurrent and convolutional models suit stations with long, stable histories, graph models earn their cost where the network is dense and spatial spillovers are strong, and attention based models suit event driven peaks that recur irregularly.(8,12,14)

This alignment between model and problem structure is more consistent across the axis than any claim that one family dominates the others.

The evidence is also unbalanced across modes. Rail networks dominate the core studies, because their gated stations produce complete and consistent validation records, whereas bus demand is noisier and its records more fragmented.

The imbalance is consequential, because conclusions drawn from metro data, where boarding is observed and schedules are rigid, may not survive the looser structure of a bus network, and the axis reads the confidence of the rail literature with this qualification in mind.(9,11)

Computational cost and reproducibility receive less attention than accuracy. Deeper and hybrid models require more data, longer training and more careful tuning, and the reviewed studies do not report these costs in a comparable form. Reproducibility is further limited because data and code are shared in only a minority of cases, which makes independent confirmation of the reported gains difficult.(23,24)

A related open question is where the gains of a complex model actually come from. Few of the reviewed studies run the ablations that would attribute an improvement to a specific component, such as the graph layer or the attention mechanism, rather than to better tuning or more data. In the absence of such attribution, the field risks accumulating architectures whose individual contributions are never isolated.(23,24)

Seventeen of the core studies sustain this axis.

Data sources and features

Smart card and automatic fare collection records remain the dominant empirical basis of the field. Aggregated into boarding, alighting or origin destination series, these records support models at the stop, station, line and network level, and the chosen aggregation is itself a design decision that conditions the achievable error.(25,26)

The granularity of this aggregation, whether five minute intervals at a single stop or daily totals across a corridor, determines both the signal that a model can learn and the noise it must absorb.

The temporal and spatial granularity at which demand is modelled is as consequential as the choice of learner. Fine temporal bins preserve signal but magnify noise, and coarse bins smooth the very peaks that operators need to see.

The reviewed studies therefore treat aggregation as a hyperparameter to be tuned rather than a default, and they construct features from calendar position, service type and network context before the model is trained.(25,27)

Observability is partial by construction. Trips without validation and corridor transfers are absent from the raw record, and alighting stops have to be inferred from travel history rather than observed directly.(26,28)

This partial observability must be handled explicitly rather than ignored, because treating an incomplete record as complete biases the demand estimates that downstream models inherit.

Origin destination estimation forms a distinct and harder subproblem within the same data regime. Station level counts are observed directly, but the flows between station pairs have to be reconstructed from entry and exit validations, from travel history or from assignment assumptions, and the resulting matrices are both sparser and noisier than the counts from which they are built.(26,28)

The distinction matters because many planning decisions, from crowding management to network redesign, require flow estimates rather than station totals.

Data quality is a second, quieter constraint. Fare policy changes, line renumbering and equipment failures alter the meaning of a series without necessarily appearing in the raw record, and a model that learns from an inconsistent series will inherit that inconsistency.(28,29)The axis therefore treats cleaning and reconciliation as part of the modelling workflow rather than as preparation that precedes it.

Open access and fare free systems, where boarding is not tied to a validation, remove the very signal on which the field depends. For such systems the evidence turns to alternative counts and to trip planner or multi source signals, and the shift changes both the coverage and the error structure of the resulting series.(27,30)

This exposes a dependence of the field on fare collection infrastructure that is easy to overlook when most studies come from gated metro networks.

The availability of long, clean histories is itself uneven. Large metro networks can accumulate years of validations, whereas smaller cities and newly opened lines begin with little or nothing, and the methods that rely on abundant data are precisely those least able to help where demand information is scarcest.

This asymmetry links the data axis to the deployment axis, because the smaller operator is the one that most needs transferable models and is least likely to have the data science capacity to build them.(25,29)

Exogenous variables provide consistent but modest gains. Weather and calendar features improve short term accuracy across bus and rail studies, whereas built environment and land use variables sharpen the spatial structure of demand where station contexts differ sharply.(26,31) Medium term studies privilege network, timetable and sociodemographic determinants when the question is route level planning rather than next hour prediction.(29)

The magnitude of the weather and calendar contribution deserves a caveat. These variables explain a real but small share of the variation once the schedule and the recent history of a station are accounted for, and their effect differs by mode and by climate. Rail demand behind a stable timetable is less weather sensitive than bus demand, and extreme events matter more than ordinary temperature swings. The axis therefore treats exogenous variables as refinements rather than as substitutes for the temporal structure that the model already captures.(26,31)

Emerging sources widen coverage beyond ticketing. Multi source streams and trip planner queries have each been shown to carry predictive signal for short term demand, at the price of irregular availability and unresolved privacy constraints.(27,30)Seven of the core studies sustain this axis.

Predictive performance and benchmarking

Across the reviewed studies, accuracy degrades as the forecast horizon lengthens, and relative errors are largest in low demand stations and off peak periods where small counts are volatile. Studies that evaluate service reliability over variable horizons document this interaction and treat the horizon as an operational choice rather than a fixed setting.(32)

The implication is that a single headline error hides the real distribution of performance, and that the stations where forecasting is hardest are also those where an agency can least afford surprises.

The metric itself shapes the conclusions. Mean absolute and root mean square errors weight stations and periods differently, percentage metrics inflate errors at low demand stops, and a model that wins on one measure can lose on another. The reviewed benchmarking work therefore reports several metrics and draws conclusions only where the ranking is stable across them.(33)

Metrics that are standard in the machine learning community do not always map onto the quantities that a transit agency optimizes, and the axis shows this mismatch to be a persistent

source of confusion.

Direct comparability is nonetheless weak. The field lacks a shared benchmark, and comparisons across identical data, horizons and error metrics are scarce, so the relative standing of two architectures is not settled by independent comparison.(33,34)

Evaluation protocols introduce a further source of divergence, because walk forward validation, random hold out splits and cross validation can leak temporal structure in different ways, and the post pandemic break makes any split across that period difficult to interpret.(34)

A subtler threat to the reported numbers is lookahead bias. When a model is tuned, or its features are standardized, using information from the evaluation period, the resulting error is optimistic in a way that survives replication only as long as the leakage is not examined.

The reviewed studies vary in how explicitly they describe the separation between training and test periods, and the axis regards transparent splitting as a precondition for taking any error figure at face value.(33,34)

The value of a forecast is not measured by its average error alone. Overprediction and underprediction carry asymmetric costs in public transport, because an overcrowded vehicle and an empty one are not symmetric failures, and the operational decision that follows a forecast determines which direction of error matters more.(32,35)

Few of the reviewed studies evaluate models against a decision aware loss, which leaves the practical value of increasingly accurate point forecasts partly unquantified.

Reporting standards would also help. A minimal disclosure of dataset size, the aggregation interval, the forecast horizon, the split rule, the metrics and the computational cost would allow readers to locate a study in the space of the field without reimplementing it, and the absence of such disclosure is one reason why the axis can report patterns but not magnitudes.(33,34)

Uncertainty quantification remains exceptional despite its value for capacity and crowding decisions. Probabilistic formulations that return distributions rather than point forecasts are beginning to appear, but no common calibration standard has been adopted.(35)Four of the core studies sustain this axis.

Deployment challenges

Accuracy in the laboratory does not guarantee operational usefulness. Dynamic services impose latency, robustness and data quality requirements that a precise offline model may not satisfy, and the reviewed work formalizes these requirements as design constraints rather than afterthoughts.(36,37)

A model that needs minutes of computation or that degrades silently when a feed fails is not deployable regardless of its benchmark score.

Real time deployment also changes the nature of the task, because the model must consume a stream, tolerate late or missing feeds and issue forecasts on a fixed cadence.(36)

The reviewed evidence indicates that these constraints are met more easily by compact models and by architectures that degrade gracefully, a property that benchmark studies do not measure.(37)

Privacy and data governance form a second deployment constraint that the methodological literature understates. High resolution demand data are personal in the aggregate, and the rules that govern their storage, linkage and reuse differ across jurisdictions, so a model that is lawful to train in one setting may not be in another.(36,37)

The deployment literature begins to acknowledge this, but governance is still treated as an external condition rather than as a design requirement.

Generalization to stations and networks with little historical data is a recurring barrier. Transfer learning, domain adaptation and meta learning mitigate this scarcity by reusing knowledge from data rich contexts, and cross city and cross mode experiments report usable accuracy at new sites, although negative transfer is typically unmeasured.(38-42)The evidence suggests that these techniques are most valuable for newly opened stations and for smaller networks that cannot accumulate long histories.

The deployment question also has a cost side that the literature underreports. A forecasting system that shaves a few percentage points of error is worth its cost only when the resulting decisions, a bus added, a service held, a crew rescheduled, are themselves valuable, and the reviewed studies do not connect forecast error to a monetary or service outcome. Without that connection, agencies cannot judge whether the marginal complexity of a deeper model pays for its maintenance, and the field is left recommending sophistication on the strength of accuracy alone.(36,37)

Interpretability has moved from an optional feature to an expected property, driven by the

need to justify dispatch and service changes. Intrinsically explainable ensembles and post hoc explanation tools are increasingly integrated into demand models, yet the explanations remain mostly descriptive and only occasionally validated with operators.(37,43)

Cross agency comparison and open data would accelerate the deployment agenda. The reviewed transfer and meta learning studies rely on data from several cities or modes, and their results depend on the willingness of operators to share series, however anonymized.(38,39)

This dependence makes open data policy a practical instrument for better forecasting rather than a transparency measure alone.

Maintenance and drift receive even less attention. The reviewed work does not follow a model after deployment to document how its accuracy changes as the network and its users evolve, and no shared practice exists for deciding when a model should be retrained.(36)

Beyond the algorithm, deployment depends on organizational conditions that the literature touches only lightly, such as access to data, the ability to maintain a pipeline and the acceptance of model outputs by planners, and these conditions are not reported alongside accuracy.(36,37) Eight of the core studies sustain this axis.

Taken together, the four axes describe a field in which methods have matured faster than data governance and operational practice. The synthesis table condenses the key findings, the evidence base and the representative studies of each axis (Table 1).

Table 1. Synthesis of the four thematic axes of machine learning for public transport demand forecasting.

Axis	Key findings	Core studies	Main limitations
Model families	Progression from statistical baselines to recurrent, convolutional, graph and transformer models; hybrid designs report the lowest errors	17	Heterogeneous datasets and metrics; inconsistent baselines
Data sources	Smart card records dominate; weather, calendar and built environment add modest gains; emerging sources widen coverage	7	Partial observability; irregular availability and privacy
Performance and benchmarking	Error grows with horizon and low demand; comparability weak; uncertainty rarely quantified	4	No shared benchmark; no calibration standard
Deployment	Latency and robustness requirements; transfer learning mitigates data scarcity; interpretability descriptive	8	Limited real deployments; negative transfer unmeasured

DISCUSSION

The practical implication of this synthesis is that an agency can obtain accurate short term demand forecasts from commodity deep learning models, but the distance between a published model and a dependable operational service remains wide. The value of machine learning is no longer the question; what is missing is the surrounding machinery of benchmarks, uncertainty and maintenance that turns accuracy into decisions.

Earlier reviews of demand determinants described a field dominated by aggregate econometric and time series models.(1)

The present evidence shows that balance has shifted, in line with broader surveys of machine learning in intelligent transport systems.(3)The recurrent architectures that now dominate descend from a long standing line of sequence modelling, and the tree ensembles used as baselines have equally deep roots.(5-7)

The convergence is not total, however, because studies diverge regarding whether complex architectures repay their cost over careful ensemble baselines, and the hybrid literature itself concedes diminishing returns.(34)

A plausible mechanism for the consistent advantage of deep models is their capacity to approximate non stationary and high dimensional patterns without pre specified structure. Demand series combine recurrent schedules, weather shocks, spatial spillovers and slow drift, and the layers of a deep network give each of these a place in the same representation.(3,5)

The same capacity explains the field's recurring weakness, because a model flexible enough to fit these patterns is also flexible enough to overfit them, which is why benchmarks and out of sample protocols matter as much as architectures.(34)

What distinguishes this synthesis from earlier reviews is the shift of emphasis from methods to deployment. Prior work documented the taxonomy of models, whereas the present reading of the evidence shows that the binding constraints have moved downstream, toward data governance, comparability and operational maintenance.(1,3)

The method question is close to settled for short term rail demand, while the questions that remain open are organizational and evaluative.(44)

The finding that smart card records dominate reflects the maturity of automatic fare collection rather than any inherent superiority of the data itself. Comparable work on ride hailing and on the spatial structure of urban demand shows the same dependence on transactional traces, which suggests that the data bottlenecks of the field are shared across modes.(45,46)

Emerging sources such as Wi Fi based mobility estimates and video based counting widen coverage, but they introduce privacy and calibration costs that the reviewed studies only begin to quantify.(47-49)

The growing demand for interpretability mirrors a wider shift in applied machine learning, where explanation is increasingly required for decisions that affect the public.(50,51)

Transit studies echo this pattern, because interpretable models of delays and of usage rates have been used to communicate findings to non technical audiences.(52,53)

The risk is that explanation remains descriptive and decoupled from the operational choices it is meant to inform.(37)

The absence of shared benchmarks and calibrated uncertainty is the most consequential weakness. Trust in forecasts for fixed guideway projects already correlates with how agencies perceive accuracy,(2)and a field that cannot compare models or report uncertainty will struggle to earn that trust. Congestion forecasting, a closely related task, has moved toward systematic comparison frameworks, and public transport demand forecasting could follow the same path.(44)

Deployment remains the least studied stage. Transfer and meta learning address data scarcity for new stations, but the reviewed work rarely measures negative transfer or long term drift, and sustained use inside an agency is seldom documented.(36,42)

Interpretability tools are descriptive, and their link to operational decisions is asserted more often than demonstrated.(37,43)

The reproducibility problems observed here are not specific to transport. The wider machine learning literature has documented how unreleased code, private datasets and selective reporting inflate the apparent performance of new methods, and the transit forecasting field reproduces that pattern in miniature.(3)

The remedy is the same: registered benchmarks, public data and code, and reporting of negative results, none of which requires new theory and all of which would raise the signal to noise ratio of the field.(34)

The unevenness of forecasting accuracy also carries an equity dimension that the methods literature seldom makes explicit. Because relative errors are largest at low demand stops and in off peak periods, the places where a model is least reliable coincide with the users and routes that agencies tend to serve with the thinnest margins. Forecasting improvements that concentrate on the busiest stations therefore reinforce existing priorities, and a deliberate effort is required to evaluate models on the routes and periods where the social stakes are highest.(29,32)

Compared with road traffic forecasting, transit demand forecasting is both easier and harder in instructive ways. It is easier because service is scheduled, so a large share of the variation is knowable in advance from the timetable, and harder because ridership responds to fare, weather and network changes that vehicle counts do not register in the same way. This duality explains why simple calendar and schedule features retain so much predictive power in transit, and why the residual that machine learning must explain is smaller but more erratic.(29,32)

A further caution concerns the geography of the evidence. A large share of the core studies was developed on metro networks in a small number of countries, and the transferability of those results to bus systems, to smaller cities and to networks with less stable demand histories remains an open question.(29,38)

Bus demand is noisier and less studied than rail demand, and the imbalance suggests that the field’s conclusions are more firmly grounded for rail than for the modes where forecasting errors may hurt most.(25,31)

The gap between research and practice is partly organizational. Academic teams are rewarded for novelty and benchmark scores, whereas agencies are rewarded for reliability and service quality, and the two incentive systems do not overlap.

The reviewed deployment studies are exceptional precisely because they were produced with or inside operators, which suggests that the most direct way to close the gap is collaborative

evaluation rather than further methodological refinement.(36,37)

For public policy, the synthesis carries a message about where investment pays off. The marginal social return of yet another architecture is low, whereas the return on shared data infrastructure, open benchmark datasets and in house data science capacity is plausibly high and falls mainly on the public sector. Agencies that treat forecasting as a service to be maintained, rather than as a one off modelling exercise, are the ones most likely to convert the field's technical progress into reliable decisions, and the reviewed evidence is consistent with that reading even though it rarely tests it directly.(2,36)

For practice, the synthesis points to three priorities: adopt simple, well benchmarked baselines before deep architectures; instrument data pipelines so that missing validations and drift are visible; and demand interval rather than point forecasts wherever capacity decisions follow. For research, the highest value contributions are shared benchmarks, calibrated probabilistic models and longitudinal studies of models in production, together with decision aware evaluation that measures the cost of errors rather than their average size. These two agendas reinforce each other, because a benchmark that no agency uses is as inert as a model that no agency can maintain.

A usable benchmark for this field need not be elaborate. It would fix a handful of public datasets, define a common train, validation and test split that respects temporal order, report a small set of error metrics together with per station and per period breakdowns, and require that every entry disclose training time and the treatment of missing data. The reviewed comparison studies already contain most of these ingredients in isolation.(33,34)

What is missing is not a technical innovation but an agreement to use them together, which is an organizational task rather than an algorithmic one.(2)

The social impact emphasis of this journal gives the equity and governance findings particular salience. A forecasting system that is accurate on average but wrong at the periphery can quietly allocate service away from the riders who have the fewest alternatives, and the deployment literature has only begun to treat such distributional effects as first class concerns rather than as side effects of accuracy.(32,36)

Reviews that focus on methods alone miss this dimension, which is why the present synthesis keeps deployment and data governance in the foreground.

Limitations

This review has four limitations, each with a corresponding remedy. The search was not exhaustive and may have missed studies indexed only in regional databases or written in languages other than English, and an openly indexed and continuously updated corpus would close this gap.

The synthesis relies on reported results rather than re analysis of primary data, so it cannot adjudicate conflicting performance claims, and shared benchmarks with public data would allow independent verification. The assignment of studies to four axes is a judgement, and individual studies cross axis boundaries, so a finer and machine readable coding scheme would reduce ambiguity. The absence of shared metrics prevented any quantitative pooling of errors, and standardized reporting would make meta analytic synthesis feasible.

CONCLUSIONS

Machine learning has consolidated itself as the default toolkit for public transport demand forecasting, with recurrent, convolutional, graph and transformer models progressively replacing statistical baselines, and hybrid designs delivering the lowest reported errors. Smart card records remain the principal data source, with weather, calendar and built environment variables adding modest but consistent gains, and emerging sources widening coverage at the cost of availability and privacy.

The main weaknesses of the field are organizational rather than methodological, because results are hard to compare for lack of shared benchmarks, uncertainty is rarely quantified, and deployment constraints such as generalization, interpretability and maintenance receive far less attention than accuracy. Closing these gaps will require less effort on new architectures and more on benchmarks, probabilistic outputs and longitudinal validation with transit agencies.

REFERENCES

1. Polat C. The Demand Determinants for Urban Public Transport Services: A Review of the Literature. J of Applied Sciences. 2012;12(12):1211-1231. doi: 10.3923/jas.2012.1211.1231
2. Voulgaris CT. Trust in forecasts? Correlates with ridership forecast accuracy for fixed-guideway transit projects. Transportation. 2020;47(5):2439-2477. doi: 10.1007/s11116-019-10024-8
3. Hassan M, Al Nafees A, Shraban SS, Paul A, Mahin HD. Application of machine learning in intelligent transport systems: a comprehensive review and bibliometric analysis. Discov Civ Eng. 2025;2(1):98. doi: 10.1007/s44290-025-00256-2
4. Baethge C, Goldbeck-Wood S, Mertens S. SANRA—a scale for the quality assessment of narrative review articles. Res Integr Peer Rev. 2019;4(1):5. doi: 10.1186/s41073-019-0064-8
5. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation. 1997;9(8):1735-1780. doi: 10.1162/neco.1997.9.8.1735
6. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32. doi: 10.1023/a:1010933404324
7. Chen T, Guestrin C. XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785-794. doi: 10.1145/2939672.2939785
8. Fontes T, Correia R, Ribeiro J, Borges JL. A Deep Learning Approach for Predicting Bus Passenger Demand Based on Weather Conditions. Transport and Telecommunication Journal. 2020;21(4):255-264. doi: 10.2478/ttj-2020-0020
9. Cao B, Li Y, Chen Y, Yang A. A CNN-LSTM Model for Short-Term Passenger Flow Forecast Considering the Built Environment in Urban Rail Transit Stations. J Transp Eng, Part A: Systems. 2024;150(11):04024072. doi: 10.1061/jtepbs.teeng-8579
10. Li S, Liang X, Zheng M, Chen J, Chen T, Guo X. How spatial features affect urban rail transit prediction accuracy: a deep learning based passenger flow prediction method. Journal of Intelligent Transportation Systems. 2024;28(6):1032-1043. doi: 10.1080/15472450.2023.2279633
11. Luo Y, Zheng J, Wang X, Tao Y, Jiang X. A Neural Network Based on Spatial Decoupling and Patterns Diverging for Urban Rail Transit Ridership Prediction. IEEE Trans Intell Transport Syst. 2023;24(12):15248-15258. doi: 10.1109/tits.2023.3308949
12. Zhang J, Mao S, Zhang S, Yin J, Yang L, Gao Z. EF-former for short-term passenger flow prediction during large-scale events in urban rail transit systems. Information Fusion. 2025;117:102916. doi: 10.1016/j.inffus.2024.102916
13. Wang T, Song J, Zhang J, Tian J, Wu J, Zheng J. Short-term metro passenger flow prediction based on hybrid spatiotemporal extraction and multi-feature fusion. Tunnelling and Underground Space Technology. 2025;159:106491. doi: 10.1016/j.tust.2025.106491
14. Baghbani A, Rahmani S, Bouguila N, Patterson Z. TMS-GNN: Traffic-aware Multistep Graph Neural Network for bus passenger flow prediction. Transportation Research Part C: Emerging Technologies. 2025;174:105107. doi: 10.1016/j.trc.2025.105107
15. Yan B, Wang Y, Huang L, Wan C, Sun W, Liu Z, et al. MSP-GATNet: a multi-scale spatiotemporal graph attention network for metro OD flow prediction. Transportmetrica A: Transport Science. 2025:1-33. doi: 10.1080/23249935.2025.2565713
16. Zhao T, Huang Z, Tu W, Biljecki F, Chen L. Developing a multiview spatiotemporal model based on deep graph neural networks to predict the travel demand by bus. International Journal of Geographical Information Science. 2023;37(7):1555-1581. doi: 10.1080/13658816.2023.2203218
17. Zhao T, Huang Z, Tu W, He B, Cao R, Cao J, et al. Coupling graph deep learning and spatial-temporal influence of built environment for short-term bus travel demand prediction. Computers, Environment and Urban Systems. 2022;94:101776. doi: 10.1016/j.compenvurbsys.2022.101776
18. Wang J, Zhang Y, Hu Y, Yin B. Metro Flow Prediction With Hierarchical Hypergraph Attention Networks. IEEE Trans Artif Intell. 2024;5(6):3012-3021. doi: 10.1109/tai.2023.3337052
19. Behroozi A, Edrisi A. Predicting travel demand of a bike sharing system using graph convolutional neural networks. Public Transp. 2025;17(1):281-317. doi: 10.1007/s12469-024-00371-w
20. Long S, Xie J, Cai M. MFMM: a passenger flow prediction model for rail transit systems considering multiple external environmental factors at stations. Transportmetrica A: Transport Science. 2025:1-33. doi: 10.1080/23249935.2025.2501203
21. Wang X, Chen D, Ye X, Zhao X, Ni S. Short-term origin-destination passenger flow forecasting in urban rail transit systems: An ensemble deep learning approach based on data augmentation. Computers & Industrial Engineering. 2025;206:111245. doi: 10.1016/j.cie.2025.111245
22. Hu S, Chen J, Zhang W, Liu G, Chang X. Graph transformer embedded deep learning for short-term passenger flow prediction in urban rail transit systems: A multi-gate mixture-of-experts model. Information Sciences. 2024;679:121095. doi: 10.1016/j.ins.2024.121095
23. Huang H, Mao J, Kang L, Lu W, Zhang S, Liu L. A dynamic graph deep learning model with multivariate empirical mode decomposition for network-wide metro passenger flow prediction. Computer-Aided Civil and Infrastructure Engineering. 2024;39(17):2596-2618. doi: 10.1111/mice.13214
24. Wu J, He D, Jin Z, Li X, Li Q, Xiang W. Learning spatial–temporal pairwise and high-order relationships for short-term passenger flow prediction in urban rail transit. Expert Systems with Applications. 2024;245:123091. doi: 10.1016/j.eswa.2023.123091
25. Liyanage S, Abduljabbar R, Dia H, Tsai PW. AI-based neural network models for bus passenger demand forecasting using smart card data. Journal of Urban Management. 2022;11(3):365-380. doi: 10.1016/j.jum.2022.05.002

26. Tang T, Liu R, Choudhury C. Incorporating weather conditions and travel history in estimating the alighting bus stops from smart card data. Sustainable Cities and Society. 2020;53:101927. doi: 10.1016/j.scs.2019.101927
27. Wang Z, Pel AJ, Verma T, Krishnakumari P, van Brakel P, van Oort N. Effectiveness of trip planner data in predicting short-term bus ridership. Transportation Research Part C: Emerging Technologies. 2022;142:103790. doi: 10.1016/j.trc.2022.103790
28. Jiang W, Ma Z, Koutsopoulos HN. Deep learning for short-term origin–destination passenger flow prediction under partial observability in urban railway systems. Neural Comput & Applic. 2022;34(6):4813-4830. doi: 10.1007/s00521-021-06669-1
29. Egu O, Bonnel P. Medium-term public transit route ridership forecasting: What, how and why? A case study in Lyon. Transport Policy. 2021;105:124-133. doi: 10.1016/j.tranpol.2021.03.002
30. Cui H, Si B, Chi D, Li Y, Li G, Chen Y. Short-term passenger flow prediction for urban rail systems: A deep learning approach utilizing multi-source big data. PLoS One. 2025;20(10):e0333094. doi: 10.1371/journal.pone.0333094
31. Farahmand ZH, Gkiotsalitis K, Geurs KT. Predicting bus ridership based on the weather conditions using deep learning algorithms. Transportation Research Interdisciplinary Perspectives. 2023;19:100833. doi: 10.1016/j.trip.2023.100833
32. Zhou T, Wu W, Peng L, Zhang M, Li Z, Xiong Y, et al. Evaluation of urban bus service reliability on variable time horizons using a hybrid deep learning method. Reliability Engineering & System Safety. 2022;217:108090. doi: 10.1016/j.ress.2021.108090
33. Zimmo I, Hörcher D, Singh R, Graham DJ. Benchmarking Travel Time and Demand Prediction Methods Using Large-scale Metro Smart Card Data. Period Polytech Transp Eng. 2023;51(4):357-374. doi: 10.3311/pptr.22252
34. Hightower A, Ziedan A, Guo J, Zhu X, Brakewood C. A comparison of time series methods for post-COVID transit ridership forecasting. Journal of Public Transportation. 2024;26:100097. doi: 10.1016/j.jpubtr.2024.100097
35. Wang Q, Wang S, Zhuang D, Koutsopoulos H, Zhao J. Uncertainty Quantification of Spatiotemporal Travel Demand With Probabilistic Graph Neural Networks. IEEE Trans Intell Transport Syst. 2024;25(8):8770-8781. doi: 10.1109/tits.2024.3367779
36. Peled I, Lee K, Jiang Y, Dauwels J, Pereira FC. On the quality requirements of demand prediction for dynamic public transport. Communications in Transportation Research. 2021;1:100008. doi: 10.1016/j.commtr.2021.100008
37. Barbareschi M, Emmanuele A, Mazzocca N, Rocco di Torrepadula F. Designing on-board explainable passenger flow prediction. Engineering Applications of Artificial Intelligence. 2025;139:109648. doi: 10.1016/j.engappai.2024.109648
38. Yin T, Nassir N, Leong J, Tanin E, Sarvi M. Transferable supervised learning model for public transport service load estimation. Transportation. 2025;52(1):29-54. doi: 10.1007/s11116-023-10411-2
39. Hua M, Pereira FC, Jiang Y, Chen X, Chen J. Transfer learning for cross-modal demand prediction of bike-share and public transit. Journal of Intelligent Transportation Systems. 2025;29(6):640-653. doi: 10.1080/15472450.2024.2371913
40. Liang Y, Huang G, Zhao Z. Cross-Mode Knowledge Adaptation for Bike Sharing Demand Prediction Using Domain-Adversarial Graph Neural Networks. IEEE Trans Intell Transport Syst. 2024;25(5):3642-3653. doi: 10.1109/tits.2023.3322717
41. Li C, Bai L, Liu W, Yao L, Waller ST. A multi-task memory network with knowledge adaptation for multimodal demand forecasting. Transportation Research Part C: Emerging Technologies. 2021;131:103352. doi: 10.1016/j.trc.2021.103352
42. Han K, Zhang J, Tian X, Li S, Zhu C. Meta-learning based passenger flow prediction for newly-operated stations. Geoinformatica. 2024;28(3):433-457. doi: 10.1007/s10707-023-00510-8
43. Yu J, Chang X, Hu S, Yin H, Wu J. Combining travel behavior in metro passenger flow prediction: A smart explainable Stacking-Catboost algorithm. Information Processing & Management. 2024;61(4):103733. doi: 10.1016/j.ipm.2024.103733
44. Attioui M, Lahby M. Congestion Forecasting Using Machine Learning Techniques: A Systematic Review. Future Transportation. 2025;5(3):76. doi: 10.3390/futuretransp5030076
45. Zhang X, Zhao X. Machine learning approach for spatial modeling of ridesourcing demand. Journal of Transport Geography. 2022;100:103310. doi: 10.1016/j.jtrangeo.2022.103310
46. Chainikov D, Zakharov D, Kozin E, Pistsov A. Studying Spatial Unevenness of Transport Demand in Cities Using Machine Learning Methods. Applied Sciences. 2024;14(8):3220. doi: 10.3390/app14083220
47. Fabre L, Bayart C, Bonnel P, Mony N. The potential of Wi-Fi data to estimate bus passenger mobility. Technological Forecasting and Social Change. 2023;192:122509. doi: 10.1016/j.techfore.2023.122509
48. McCarthy C, Ghaderi H, Martí F, Jayaraman P, Dia H. Video-based automatic people counting for public transport: On-bus versus off-bus deployment. Computers in Industry. 2025;164:104195. doi: 10.1016/j.compind.2024.104195
49. Jovanović B, Shabanaj K, Ševrović M. Conceptual Model for Determining the Statistical Significance of Predictive Indicators for Bus Transit Demand Forecasting. Sustainability. 2022;15(1):749. doi: 10.3390/su15010749
50. Sariyer G, Mangla SK, Sozen ME, Li G, Kazancoglu Y. Leveraging explainable artificial intelligence in understanding public transportation usage rates for sustainable development. Omega. 2024;127:103105. doi: 10.1016/j.omega.2024.103105
51. Tang T, Zhang J, Chen S, Mo P, Pei M, Tang TQ. Deciphering the pulse of the city: An exploration of the natural features of metro passenger flow using XAI. Computers & Industrial Engineering. 2025;204:111097. doi: 10.1016/j.cie.2025.111097

52. Lviv Polytechnic National University, S. Bandera Street, 12, Lviv, 79013, Ukraine, Matseliukh Y, Lytvyn V, Hu Z, Bublyk M. Predictive Modelling and Factor Analysis of Public Transport Delays in Smart City Using Interpretable Machine Learning. IJITCS. 2025;17(6):1-28. doi: 10.5815/ijitcs.2025.06.01
53. Sarhani M, Nourmohammadzadeh A, Voß S, EL Amrani M. Predicting and analyzing ferry transit delays using open data and machine learning. Journal of Public Transportation. 2025;27:100124. doi: 10.1016/j.jpubtr.2025.100124

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Carlos Ignacio Ruiz Maldonado.
Writing–review and editing: Carlos Ignacio Ruiz Maldonado.