The Role of Spatial Features and Adjacency in Data-Driven Short-Term Prediction of Trip Production: An Exploratory Study in The Netherlands

Zahra Eftekhar, Saman Behrouzi, Panchamy Krishnakumari, Adam Pel, and Hans van Lint

***Abstract*—Large-scale prediction of trip production is essential for origin–destination (OD) demand estimation and prediction. One of the main challenges in predicting trip production patterns lies in addressing spatial-temporal correlations and variations. Whereas many studies focus on temporal correlations, very few consider spatial adjacency between traffic analysis zones (TAZ) as explanatory variables. This research proposes a method that integrates a graph convolutional neural network (GCN) into a long short-term memory network (LSTM) to do exactly that. By introducing a nationwide graph that encodes the adjacency of TAZs, spatial heterogeneity is considered in the prediction process, and a single prediction model is trained for the entire network, thereby avoiding the need to train multiple separate models and potentially reducing overall training overhead, while increasing the prediction accuracy. Moreover, with this model, we investigate the effect of spatial scale on spatial uncertainty and prediction accuracy and analyze prediction errors, residual patterns, and their associations with socio-spatial features at different spatial scales. The findings of this research have important implications for improving OD demand prediction models and provide valuable insights into the role of spatial scale and socio-spatial features in travel demand prediction.**

***Index Terms*—Trip production, demand prediction, spatial scale, spatial-temporal pattern, graph convolution, residual analysis.**

I. Introduction

A. Background

Short-term travel demand prediction is crucial for effective policy adjustments, management, and operations in various domains [1], [2]. A critical aspect of demand prediction is forecasting the number of trips originating from a specific location, referred to as *trip production*. Although accurate trip production prediction is vital for estimating and predicting origin-destination (OD) demand, it remains a challenging task due to its complex spatial-temporal dependence and heterogeneity [3]. In this paper, "trip production" denotes the number of outgoing trips from a particular location or zone i.

Received 4 October 2023; revised 8 May 2024, 7 December 2024, and 21 June 2025; accepted 28 July 2025. This work was supported by the NWO/TTW Project MiRRORS under Grant 16270. The Associate Editor for this article was R. Soosaimarian Peter Raj. *(Corresponding author: Zahra Eftekhar.)*

The authors are with the Transport and Planning department, Faculty of Civil Engineering and Geosciences, Delft University of Technology, 2628 CN Delft, The Netherlands (e-mail: z.eftekhar-1@tudelft.nl).

Digital Object Identifier 10.1109/TITS.2025.3610652

Predicting trip production is challenging due to the heterogeneity caused by the diverse characteristics of traffic analysis zones (TAZs) [4]. **Spatiotemporal heterogeneity** refers to the variations in the correlations or distributions of variables across different geographical regions and time intervals. Spatiotemporal heterogeneity in demand implies that the effects of land-use properties on travel demand patterns may not be consistent across different areas or periods. This heterogeneity within one TAZ and across TAZs originates from a range of factors, including diverse urbanization levels, population demographics and lifestyles, economic activities, transportation accessibility, and resource distribution across areas [5]. When a single TAZ encompasses a mixture of (diverse) characteristics, it can lead to spatial heterogeneity in trip production, introducing uncertainty in predicting trip production patterns. For example, trip production in certain parts of a TAZ may peak in the afternoon, while in others, it may peak in the morning, and in some cases, it may follow a completely different pattern [6]. Considering these challenges, this paper aims to develop data-driven prediction models that explicitly account for spatial-temporal heterogeneity, investigate the effect of spatial scale on prediction accuracy, and examine the association between prediction errors, residual patterns, and socio-spatial features at different spatial scales. Understanding these relationships can significantly enhance the quality of trip production predictions and contribute to more effective urban planning and transportation policy decisions.

In the temporal dimension of demand data, various patterns, such as periodicity, linear and non-linear trends, and holiday effects, can be observed, significantly impacting the quality of predictions. This temporal variability is particularly substantial in urban areas, where morning and afternoon peaks contribute to nearly 50% of daily travel production [7]. This variability not only exacerbates the spatial heterogeneity described earlier but also has its own effects. The coarser one discretizes the periods over which production is predicted, the larger the prediction errors may be, depending on both this temporal variability *and* the chosen period boundaries.

Numerous researchers highlight that neglecting this combined spatiotemporal heterogeneity results in higher prediction errors and that addressing spatial-temporal heterogeneity is

essential for understanding and predicting travel production (e.g., [8], [9], [10]).

Our literature review reveals a lack of data-driven models that consider these factors. Consequently, to inspire potential approaches, we first provide a brief overview of data-driven prediction methods applied to traffic data (speed and flow) to address these challenges.

B. Related Work

Various studies have investigated the temporal correlation of traffic data using diverse methods tailored to different application scenarios. Traffic forecasting approaches can be broadly divided into two categories: traditional statistical methods and deep-learning-based methods.

Traditional statistical methods include linear regression models [11], Kalman filtering [12], autoregressive integrated moving average (ARIMA) [13], K-nearest neighbor (KNN) [14], least squares support vector machines (LS-SVMs) [15], particle filter [16], hidden Markov model [17], and Gaussian process [18]. These methods often rely on strict mathematical deductions and well-defined physical meanings, limiting their applicability to less complex traffic conditions and/or smaller traffic data sets. Furthermore, most traditional methods assume linearity (in their parameters) and consider stationary underlying processes generating the data, making them less suitable for non-linear dynamics and non-recurrent situations that are characteristic of traffic demand dynamics.

Deep-learning-based methods, the second category, include convolutional neural networks (CNNs) and recurrent neural networks (RNNs). CNNs, along with their deeper architecture (e.g. ResNet [19]), are typically employed for spatial structure learning, while RNNs (e.g., LSTM [20] and gated recurrent unit (GRU) [21]) are widely used for temporal and sequential learning. For instance, ST-ResNet [22] employs a residual network for spatial correlation learning and LSTM for modeling time-series data. However, ST-ResNet and subsequent works [22], [23], [24] mainly focus on "coarse-level" citywide traffic flow estimation using taxi or bicycle data.

Deep-learning-based methods often require large datasets and high computational capacity due to the large number of parameters that need to be trained. Moreover, for large-scale trip production, separate models must be trained for each spatial area, necessitating extensive computation time and resources. When developing transportation models, it is essential to find a balance between the complexity of the model and the available computing resources [25]. This is crucial because model acceptability often depends on computation time, which must be minimized without compromising accuracy. Additionally, most non-parametric methods mentioned above focus solely on temporal correlations without reflecting spatial heterogeneity.

Recent advancements in traffic prediction models have utilized deep learning techniques to capture spatial-temporal dependencies. For example, Li et al. [26] proposed a Graph WaveNet model for traffic forecasting, which integrates graph convolutional networks with temporal convolutional networks. Similarly, Guo et al. [27] developed an attention-based spatial-temporal graph convolutional network (ASTGCN) for traffic flow prediction. These studies highlight the ongoing development of models that effectively capture complex spatial-temporal relationships.

Starting with travel demand prediction, the comparison study of [28] on short-term traffic demand prediction methods reveals no singularly superior model across several time-series models, machine learning models, and three deep learning models across multiple datasets. On the whole, statistical approaches were less effective, in contrast, machine learning techniques and LSTM-based neural networks exhibited enhanced outcomes as measured by the SMAPE metric. Methods employing LSTM with one-hot encoding and LSTM with embedding techniques achieved commendable performance regarding RMSE. Generally, the LSTM-Neural Network model surpassed alternative models in comparative analyses over diverse geographical units.

In another demand prediction study, hierarchical reconciliation approaches have been explored to enhance deep-learning methods, as demonstrated by [29], who emphasized the necessity of incorporating error analysis to maintain forecast accuracy within a feasible solution space, underscoring the flexibility and applicability of these methods across various scenarios of area-based traffic demand prediction.

In the domain of traffic flow prediction, the spatial-temporal convolutional model introduced by [30] stands out for its application on urban crowd density prediction using mobile-phone signaling data, demonstrating the potential of deep learning models to handle irregular-shaped divisions and capture the intricate spatial and temporal dependencies inherent in traffic prediction tasks. This approach aligns with the findings of [31], who utilized deep learning to predict daily usage of Bike Sharing Systems (BSS), indicating a shift towards data-driven models that leverage deep learning for enhanced predictive capabilities.

Furthermore, the research by [32] into urban rail transit (URT) underscores the impact of spatial features on prediction accuracy, presenting a deep learning-based passenger flow prediction method that incorporates spatial characteristics such as land use, regional location, and intermodal access. This methodological shift towards acknowledging the role of spatial features in prediction models offers a nuanced understanding of travel behavior and demand, facilitating more accurate forecasts.

DeepSTCL, a deep spatio-temporal ConvLSTM framework for travel demand prediction developed by [33], represents another stride forward, treating historical travel data akin to a video stream to predict future demand. This innovative approach underscores the potential of deep learning in tapping into spatio-temporal features for travel demand forecasting, achieving high accuracy and efficiency. In conclusion, the integration of deep learning into travel demand and traffic flow prediction models has facilitated a deeper understanding of complex spatial and temporal dynamics, significantly enhancing predictive accuracy.

Recently, researchers have explored graph-theoretic approaches to model traffic data collected from road sensors. The spatial correlations between traffic sensors are structured as a directed graph, with nodes representing sensors and edge

weights indicating proximity between sensor pairs measured by road network distance. Given that sensor networks are naturally organized as graphs, recent advances in graph neural networks [34], particularly graph convolution networks [35], have inspired several graph-based traffic prediction models [24], [27], [36], [37], [38]. For example, [39], [40] propose a method integrating graph convolutional neural network (GCN) into the LSTM, addressing both spatiotemporal dependencies and heterogeneity.

The study by [41] presents an interesting short-term traffic flow prediction approach. This research acknowledges that most prediction models face limitations in predicting local variation patterns and handling dynamic traffic due to their reliance on training data. To address this issue, they propose a two-step approach that combines a baseline model constructed from historical data with a time-varying Vasicek (EV) model better to capture real-time variations in traffic flow during each day. What makes this study particularly interesting is its use of residuals, which are the differences between actual and predicted values, as a feature to improve traffic flow prediction. While the study does not explicitly identify and evaluate the patterns of residuals, its objective is to enhance prediction performance by considering daily uncertainty impacts by applying the EV model. This approach hints that evaluating prediction residuals can provide valuable insights into prediction errors and the factors that trigger them.

Several recent studies have proposed novel models that integrate spatial adjacency into traffic prediction. For example, Zhao et al. [42] proposed the Temporal Graph Convolutional Network (T-GCN), which combines GCN and gated recurrent units (GRU) for traffic forecasting. Similarly, Yu et al. [36] introduced a spatio-temporal graph convolutional network (STGCN) for traffic forecasting. These models demonstrate the effectiveness of incorporating spatial adjacency and temporal dynamics.

Beyond T-GCN [42] and STGCN [36], other approaches have introduced more complex frameworks to improve GCN-based traffic prediction. These include models that leverage dynamic graphs to capture evolving network topologies, transformer-based GCN architectures that integrate attention mechanisms, and hierarchical GCNs (HGCNs) that exploit spectral clustering of regions. Such methods can potentially capture evolving spatial-temporal dependencies and multi-level spatial structures, leading to further enhancements in prediction performance [43], [44], [45]. Our work builds upon this foundation by integrating GCN with LSTM to capture spatial-temporal correlations in trip production prediction. Building upon the above body of work, we next outline our own GCN-augmented approach.

Inspired by the studies discussed, we propose using the LSTM approach as our benchmark model. To account for the spatial adjacency often overlooked in conventional forecasting methods, we then incorporate a GCN within the LSTM framework. Adjacency, in the context of TAZs, refers to the geographical proximity between zones. Two zones are considered adjacent if they share a boundary or are close enough to influence each other's travel patterns. This relationship is captured through an adjacency matrix, which encodes the connections among zones.

This deliberate integration aims to evaluate the incremental improvements in prediction accuracy afforded by considering the adjacency of zones. The LSTM model establishes a solid baseline, enabling us to highlight the specific advantages of integrating spatial characteristics through the GCN. Our approach does not purport to introduce a novel modeling technique per se; rather, it illustrates the potential for enhancing prediction accuracy by integrating spatial adjacency with established predictive models. By applying this LSTM+GCN framework to the context of trip production prediction, our study underscores the significant role of spatial adjacency. While numerous studies have explored OD matrix prediction utilizing either spatial or temporal prediction models separately, our study advances this by integrating both spatial and temporal dimensions using a GCN combined with LSTM. This integration allows the model to simultaneously consider the spatial adjacency of TAZs and their temporal trip production characteristics, enhancing the accuracy and reliability of predictions over traditional models that consider these aspects in isolation.

C. Research Objectives and Contributions

Drawing on the existing literature on traffic data, this research delves into the emerging role of graph knowledge in the context of travel demand prediction. The novelty of this study lies in *combining* a single, nation-scale graph representation with a multi-scale evaluation protocol and a residual-driven socio-spatial analysis – a combination that, to the best of our knowledge, has not been reported in the OD-demand literature. We propose a method that integrates GCN into the LSTM and aims to achieve three main objectives. First, we seek to determine the extent to which computation time and prediction accuracy are impacted by incorporating the adjacency of traffic analysis zones into the prediction model. Second, we explore the influence of spatial scale on spatial uncertainty (associated with spatial heterogeneity), as this aspect has remained relatively unexplored in the demand prediction field. We compare the prediction accuracy changes across multiple administrative spatial scales when using the same model. Third, we examine prediction errors by identifying and evaluating dominant patterns of trip production prediction residuals. Lastly, we pinpoint the most critical demographic and land-use features of TAZs contributing to their associated residual clusters and compare them across different spatial scales.

Our choice to focus on the LSTM model and its integration with GCN (LSTM+GCN) was deliberate and aimed at highlighting the specific benefits of considering the adjacency of zones in forecasting models. The LSTM model served as a robust benchmark to establish a baseline for predictive performance. By integrating GCN, we sought to underscore how spatial characteristics could enhance prediction accuracy beyond the baseline. This approach was not intended to claim broad novelty in modeling techniques but rather to

demonstrate the incremental gains in prediction accuracy that can be achieved by incorporating spatial adjacency into well-established models. Therefore, our study contributes to the field by applying the LSTM+GCN framework to the specific context of trip production prediction, emphasizing the role of spatial adjacency.

The integration of GCN with LSTM in a unified framework allows for the simultaneous modeling of both spatial relationships and temporal dependencies. This dual capability is particularly advantageous in our context for several reasons. By capturing the spatial adjacency of TAZs through GCNs and the temporal trends through LSTMs, the model can make more informed, context-aware predictions than would be possible by considering either aspect in isolation. This approach allows the model to utilize the full spectrum of available data—spatial configurations and temporal sequences—thereby maximizing the insights gained from the data and improving prediction robustness. Incorporating both spatial and temporal data reflects real-world conditions more accurately, making the model's outputs more reliable and applicable for planning and operational purposes.

Our proposed method offers several key contributions to the field of travel demand prediction:

1) By employing a nationwide graph to integrate spatial adjacency into an LSTM-based model, our approach simultaneously enables large-scale, national-level trip production predictions and forgoes the need to train multiple separate models for each region or scale. This unified framework not only broadens the spatial scope of demand predictions but also potentially streamlines computational processes and resource allocation.
2) Our model's incorporation of spatial adjacency information directly addresses the spatial heterogeneity inherent in TAZs, thereby refining the accuracy of trip production forecasts.
3) Through a nuanced examination of how spatial scale influences the predictability and uncertainty of trip production, our study sheds light on crucial considerations for transport modelers and policymakers, particularly regarding the implications of spatial resolution on demand forecasts.
4) Our analysis extends beyond mere prediction errors to systematically dissect and interpret the prevailing patterns in trip production prediction residuals. By identifying the demographic and land-use characteristics most instrumental in shaping these patterns across various TAZs, our research deepens the collective understanding of trip production variability. Such insights pave the way for the formulation of more precise and robust demand prediction models tailored to different spatial scales.

Beyond the methodological innovation, this paper significantly contributes to the analysis of prediction results at multiple spatial scales, examining the association between prediction accuracy (residual patterns) and the types of built environment variables. This research advances travel demand models by systematically integrating socio-spatial characteristics, such as land use, points of interest, and demographics, identified as key influencers of travel demand across various scales. This allows for the creation of more effective models, tailored to the actual needs of society. By refining models to consider these nuanced socio-spatial dynamics, we enhance their predictive accuracy and applicability, supporting the development of responsive and equitable transportation systems. This approach not only improves model accuracy but also ensures transport planning and policies are based on a comprehensive understanding of the factors driving travel demand, leading to more targeted and effective transport planning and policy-making.

In this study, we use processed aggregated derivatives of GSM traces in the form of motor-vehicles OD matrices of the Netherlands in March 2017. This implies the scope of this work pertains to motor-vehicle trip production patterns.

The remainder of this paper is structured as follows: Section II delineates the research data and the methodology employed. Section III presents and discusses the developed trip production prediction models, elucidates the impact of spatial scale on spatial uncertainty and prediction accuracy, investigates prediction errors and residual patterns, and identifies the most pertinent socio-spatial features contributing to residual patterns across various spatial scales. Finally, Section IV offers a conclusion, summarizing the key findings and their implications.

II. Methodology

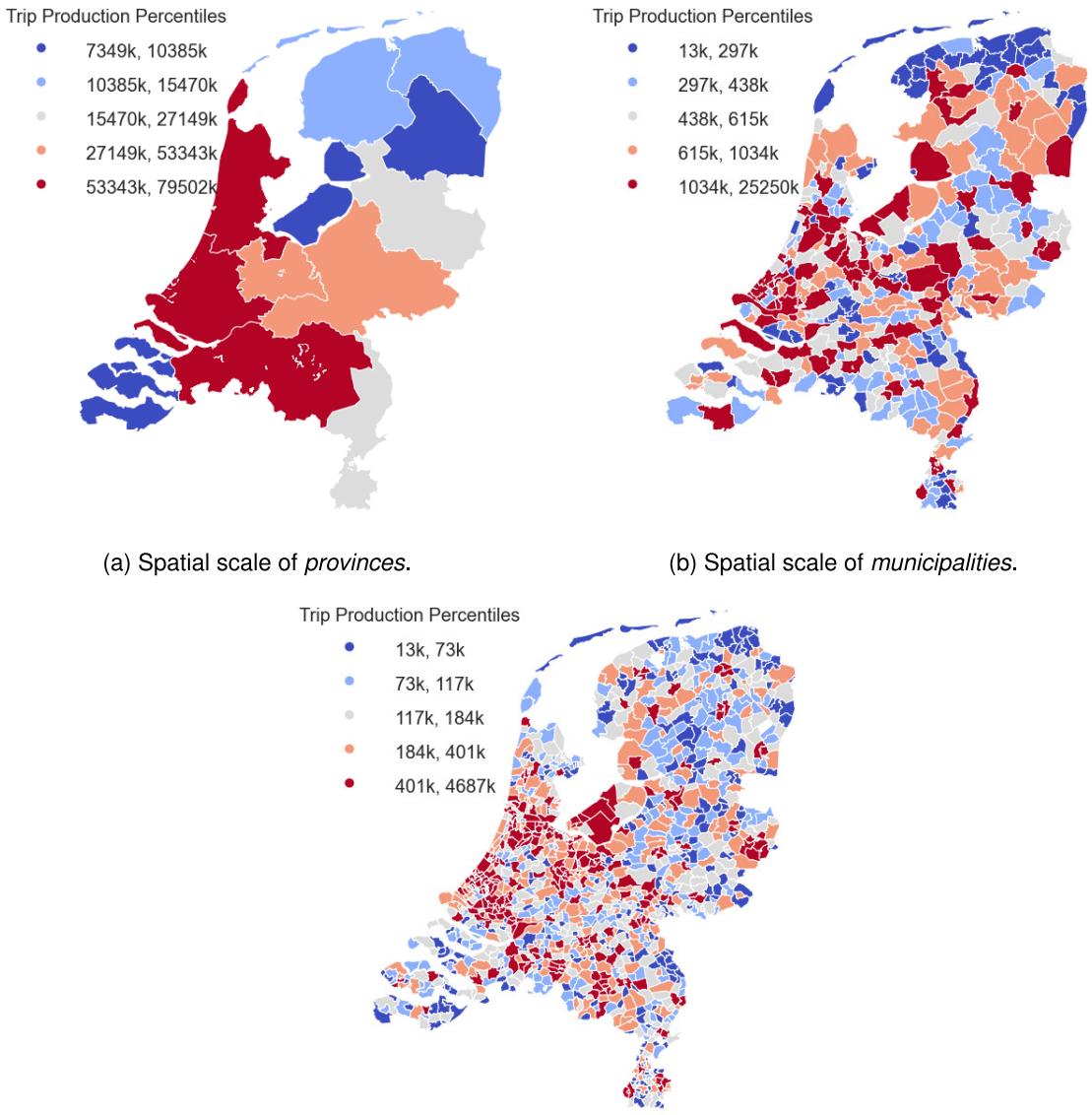
A. Trip Production and Socio-Spatial Data

This study utilizes hourly trip production data for the entire Netherlands during March 2017. The data is aggregated over three spatial scales: *provinces*, *municipalities*, and *4-digit postal code zones*, resulting in 12, 390, and 1243 Traffic Analysis Zones (TAZs) throughout the Netherlands.

Trip production for TAZ i refers to the number of inter-zonal motor-vehicle trips originating from i. Trip production values are derived from the GSM traces of Dutch telecommunications company Vodafone, which holds a market share of approximately one-third of the Dutch population. Due to privacy concerns related to raw mobile phone data, another company processed the data. As a result, this study utilizes origin-destination (OD) matrices of motor vehicles based on TAZs in the Netherlands rather than mobile phone traces. These OD matrices have been scaled up to account for the entire Dutch population. For more details on the scaling procedure, refer to [46]. All hours—including those coinciding with national holidays or local events that yield atypically high volumes—were retained; no extra temporal filtering, capping, or imputation was applied, as our objective was to evaluate the models under full real-world variability.

These OD matrices are pre-processed and reshaped to derive a vector of hourly trip production values per TAZ for each of the three spatial scales studied here in this paper. The dataset contains over 365 million produced trips. Figure 1 displays the total monthly trip production per TAZ for the three spatial scales under study.

It is important to note that the trip production data is derived from GSM traces of Vodafone users, representing



(a) Spatial scale of *provinces*.

(b) Spatial scale of *municipalities*.

(c) Spatial scale of *4-digit postal codes*.

Fig. 1. Total monthly trip production per TAZ for the three studied spatial scales.

approximately one-third of the Dutch population. While this provides substantial coverage, there is a potential risk of user selection bias or over-representation of travel patterns from specific demographic groups. The data may not fully capture the travel behavior of non-Vodafone users or certain socioeconomic segments, such as elderly populations or individuals with limited access to mobile phones. This limitation is acknowledged, and the results should be interpreted with consideration of this potential bias.

The Central Bureau of Statistics (CBS) of the Netherlands provided socio-spatial data containing approximately 140 demographic and land-use variables for each TAZ [47] at each of the three aforementioned spatial scales. CBS collects, edits, and publishes national statistics based on registers, surveys, and interviews. The variables used in our analysis correspond to data from 2017, aligning with the trip production data.

B. Prediction of Trip Production With LSTM

The trip production data exhibit strong seasonality effects and dynamic trends within recent time frames. Unlike a regular RNN, which struggles to establish long-term dependencies, the LSTM overcomes the vanishing gradient problem, enabling it to capture both short- and long-term temporal patterns in a time series. Consequently, we employ an LSTM to investigate the periodic (i.e., daily) dependencies and recent dynamic trends in trip production. We refer readers to [20] for more details on the LSTM architecture. Figure 2 illustrates the experimental framework for predicting trip production using LSTM. The model takes as input the time series signals $X_{i,s}$ representing the trip production data for each TAZ i within the spatial scale s. The data are then fed into the LSTM network, which consists of multiple hidden layers. Each LSTM layer is composed of three gates: input, forget, and output gates. The gates function as follows:

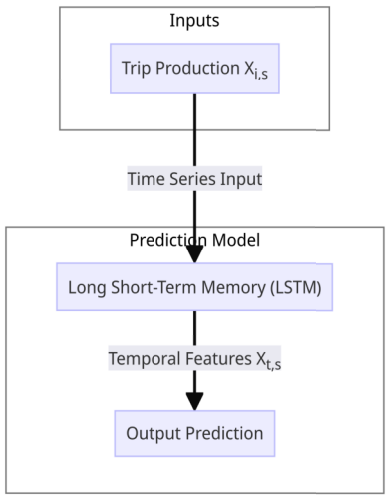


Fig. 2. Experimental framework for the trip production using LSTM.

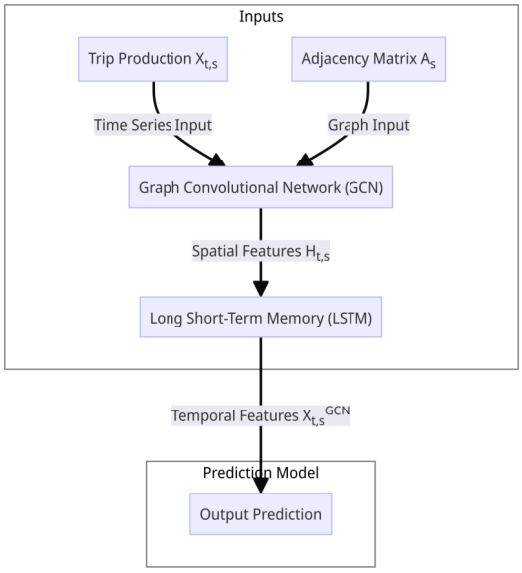


Fig. 3. Experimental framework for the trip production using GCN+LSTM.

- **Input Gate**: Determines which input information is allowed into the memory cell.
- **Forget Gate**: Controls which historical data should be retained in memory.
- **Output Gate**: Produces the final output of the memory cell and contributes a portion of the input information for the subsequent cell.

The LSTM network learns the temporal patterns in trip production data, capturing both short-term and long-term dependencies through these gate mechanisms. The output of the LSTM layer is passed to a fully connected layer, which predicts future trip production values based on the learned temporal features. In this framework, the trip production signal $X_{i,s}$ is normalized to ensure comparability across different spatial scales and TAZs, allowing the model to generalize better across regions.

C. Prediction of Trip Production With LSTM+GCN

Considering spatial correlation and heterogeneity between TAZs potentially improves trip production models. To address this, we integrate a Graph Convolutional Network (GCN) into the LSTM. We refer readers to [39] and [48] for more details. A GCN is a spatial feature extraction model applicable to any topological structure graph. To compute the GCN at time t for an M-feature matrix $X_t \in \mathbb{R}^{N\times M}$, we first generate an undirected graph $G = (Z, E, A)$, where the nodes Z represent the TAZs, and $|Z|$ equals the number of TAZs, N, in each spatial scale.

Considering spatial correlation and heterogeneity between TAZs potentially improves trip production models. To address this, we integrate a Graph Convolutional Network (GCN) into the LSTM. We refer readers to [39] and [48] for more details. A GCN is a spatial feature extraction model applicable to any topological structure graph. To compute the GCN at time t for an M-feature matrix $X_t \in R^{N\times M}$, we first generate an undirected graph $G = (Z, E, A)$, where the nodes Z represent the TAZs, and $|Z|$ equals the number of TAZs, N, in each spatial scale.

Nodes (Z): Each node represents a TAZ at a specific spatial scale. The total number of nodes $|Z| = N$ corresponds to the number of TAZs at that scale.

Edges (E): The edges represent the connections between neighboring TAZs. Two TAZs are considered neighbors if they share a common boundary. We use shapefiles of administrative boundaries and a spatial join operation to determine the neighbors.

The adjacency matrix $A \in \mathbb{R}^{N\times N}$ encodes the spatial relationships between TAZs, with A_{ij} defined as:

$$A_{ij} = \begin{cases} 1 & \text{if TAZs } i\text{and } j\text{are adjacent} \\ & \text{(i.e., share a common boundary),} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The GCN layer operates on the graph structure defined by the adjacency matrix A, aggregating information from neighboring TAZs to learn spatial features. The graph convolution operation is defined as:

$$H = \sigma(\tilde{A}XW), \quad (2)$$

where $\tilde{A} = D^{-1/2}(A + I)D^{-1/2}$ is the normalized adjacency matrix with added self-connections, D is the degree matrix of $A+I$, I is the identity matrix, X is the input feature matrix (trip production data), W is a learnable weight matrix, and σ is an activation function such as ReLU. This operation allows each TAZ to aggregate information from its neighbors, effectively capturing spatial dependencies.

Figure 3 provides a detailed schematic diagram of the proposed model framework, illustrating the integration of GCN and LSTM layers for capturing spatial and temporal correlations.

We use the graph convolution operation to extract spatial features of the mobility network and the LSTM to extract temporal features of the signal. The LSTM input consists of the convolutional graph features concatenated with the original signals. In other words, for each spatial scale, the graph signals

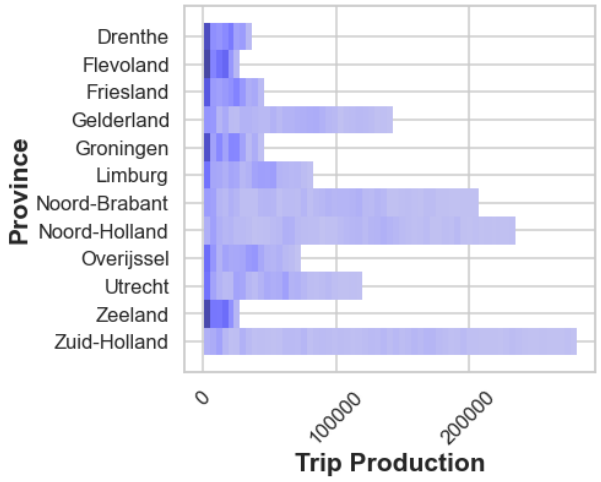


Fig. 4. Trip production values among *provinces* during March 2017.

TABLE I
LSTM AND GCN+LSTM HYPERPARAMETERS

	LSTM	GCN+LSTM
GC layers	na	2
GC layers sizes	na	16, 10
GC activations	na	relu
GC dropout	na	0.5
GC optimizer	na	Adam
GC learning rate	na	0.01
GC weight decay	na	5e-4
LSTM layers	3	3
LSTM layers sizes	128, 256, 128	128, 256, 128
LSTM activations	relu	relu
LSTM dropout	0.2	0.2
LSTM optimizer	Adam	Adam
LSTM learning rate	0.001	0.001
Fully Connected Layer	Linear	Linear
Batch size	128	128
Number of epochs	100	100
Early stopping patience	10 epochs	10 epochs
Loss function	MSE	MSE

$X_{t,s}$ at time t, along with the adjacency matrix A_s, are used to compute the spatial features $H_{t,s}$.

As depicted in Figure 3, the spatial features $H_{t,s}$ obtained from the GCN layer are concatenated with the original graph signals $X_{t,s}$ to form the input $X^{GCN}_{t,s}$ for the LSTM layer. This concatenation ensures that both the spatial information from neighboring TAZs and the temporal information from the original signals are jointly utilized by the LSTM to make predictions. Therefore, the input to the LSTM layer is:

$$X^{GCN}_{t,s} = [X_{t,s}; H_{t,s}],$$

where $[\cdot;\cdot]$ denotes concatenation along the feature dimension.

Lastly, the LSTM output serves as the input for a fully connected layer that predicts the signal for the desired time. The process can be summarized as follows:

- **Spatial Feature Extraction:** Apply the GCN layer to learn spatial features ($H_{t,s}$) from the graph signals ($X_{t,s}$) and adjacency matrix (A_s).
- **Temporal Feature Extraction:** Concatenate the spatial features ($H_{t,s}$) with the original graph signals ($X_{t,s}$) to form the input ($X^{GCN}_{t,s}$) for the LSTM layer.
- **Trip Production Prediction:** Use the LSTM layer to predict future trip production values.

For further details on the architecture and implementation of our model, including the code and data used, we have published our code repository and dataset. The code repository can be found at [49], and the dataset is available at [50].

Thus, for each spatial scale, the graph signals ($X_{t,s}$) at time (t) are concatenated with the spatial features ($H_{t,s}$) to form the input for the LSTM ($X^{GCN}_{t,s}$). The LSTM output then serves as the input for a fully connected layer that predicts the signal for the desired time.

D. Experimental Design

We initially normalize the values for each TAZ to enable a fast and stable pattern comparison across various TAZs. For instance, Figure 4 shows the 2-D histogram of trip production for all the *provinces* in the Netherlands. As shown, the value ranges for different TAZs vary significantly. We can focus more on recognizing the patterns by normalizing the values in each TAZ. We later reverse the values back to their original range to evaluate prediction accuracy. We applied the Min-Max Scaling technique to normalize each production value x, i.e.,

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3}$$

where $x_{normalized}$ is the normalized value, x_{min} and x_{max} are the minimum and maximum production values in the (month-long) time series for that particular TAZ. Consequently, the resulting normalized values range between 0 and 1. We then use the two proposed methods, LSTM and LSTM+GCN, to predict trip productions in each spatial scale. Table I provides a summary of the model parameters and hyperparameters used in this study.

We employ two evaluation metrics at each prediction step to assess our model's performance: Mean Square Error (MSE) and Mean Absolute Percentage Error (MAPE). Moreover, we compare the accuracy of the two proposed models across the studied spatial scales using the same metrics. Furthermore, we report the computation time for each model to compare the required computation capacity.

In our analysis, the selection of Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE) as pivotal metrics was guided by their respective abilities to elucidate distinct aspects of prediction. By harnessing both MSE and MAPE, our study captures a holistic view of the prediction models' performance—highlighting not just the magnitude of prediction errors but also their proportional significance relative to actual trip patterns.

MSE was chosen for its capacity to underscore and amplify the impact of larger errors in prediction. By squaring the differences between the predicted and the actual values, MSE penalizes more significant errors, thus sensitizing us to models that may be prone to occasional but substantial inaccuracies. This metric is especially critical in the prediction of trip production, where outsized errors can denote critical lapses

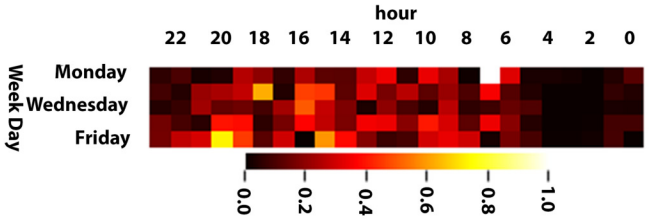


Fig. 5. An example heatmap of trip production residual of a random TAZ.

in predicting peak travel demands, a vital consideration for transportation system planning and management.

MAPE, on the other hand, offers a normative perspective by expressing prediction errors as a percentage of the actual values. This normalization allows for a more intuitive comprehension of the model's accuracy, independent of the scale of the data. MAPE's percentage-based nature renders it particularly insightful for comparative analyses across regions or time periods with varying levels of trip production. It enables us to discern the relative predictive accuracy in a manner that is agnostic to the actual trip production magnitude, thereby facilitating equitable benchmarking across diverse scenarios. Please note that MSE is also used as the training loss function for both models. MAPE is not a training loss but an evaluation metric used to compare final prediction performance.

We conducted manual hyperparameter tuning to determine the optimal configurations for the GCN and LSTM models. This involved testing various combinations of the number of layers (1 to 3), hidden units (64, 128, 256), learning rates (0.001, 0.005, 0.01), and dropout rates (0.2, 0.5). While an exhaustive search was not possible due to computational limitations, we selected the configurations that yielded the best validation performance, balancing model complexity and accuracy. Table I displays the settings we used for the implemented models in this study.

The experiments were conducted on a workstation with an Intel Core i7-9700K CPU @ 3.60GHz, 32 GB RAM, and an NVIDIA GeForce RTX 2080 Ti GPU. The models were implemented using Python 3.8 and PyTorch 1.7.1. All computations were performed locally, and the specifications are provided to ensure transparency and reproducibility.

E. Residual Analysis

To investigate prediction errors, we analyze the residual heatmap of hourly trip production prediction per Traffic Analysis Zone (TAZ). First, we predict trip production and compute the residual as the absolute difference between the predicted and actual values. To concentrate on discerning patterns, we normalize the residuals per TAZ using Min-Max Scaling. Consequently, each zone has a heatmap of normalized residual values, with each cell ranging between 0 and 1. Each heatmap's horizontal and vertical axes represent the days of the prediction horizon and the hours of the day, respectively. As an example, Figure 5 shows the heatmap of trip production residual for a randomly selected TAZ. This representation enables us to observe the temporal patterns of residuals within a day (by comparing across columns) and between days (by comparing across rows). The heatmaps for each under-study spatial scale serve as the foundation for subsequent analyses. To identify patterns of residuals, we cluster heatmaps of TAZs per spatial scale based on temporal similarity using K-means clustering. Owing to its straightforward application and interpretability, which is crucial in exploratory analyses for generating clear, understandable results, the K-means clustering method is a widely employed algorithm for clustering analysis [51], [52], [53], [54], [55]. The method aims to partition the N-dimensional dataset of M points (heatmaps) into K clusters, minimizing the sum of the pairwise Euclidean distance between the points in each cluster [56].

To identify patterns of residuals, we employed K-means clustering due to its simplicity, efficiency, and interpretability, which are crucial for exploratory analyses. K-means is a widely used algorithm that partitions the data into K clusters by minimizing the within-cluster sum of squares [57]. While other clustering methods such as hierarchical clustering or DBSCAN could be considered, K-means provides a straightforward approach that aligns well with the objectives of our study. Exploring other clustering methods is an avenue for future research.

The clustering process comprises two primary steps:

- **Assignment:** Assigning each point to its closest centroid, mathematically referring to the partitioning of the points to the Voronoi diagram [58] generated by the centroids.
- **Update:** Updating each cluster center to be the average of all points contained within them.

As the K-means method requires exogenous determination of the number of clusters, we must justify our choice. Since we cannot a priori determine the number of existing patterns, this is an unsupervised learning problem without "true labels" or ground truth. Therefore, the optimal number of clusters is determined by assessing the (dis)similarity within and between clusters for various values of K. We use the Silhouette index [59] to measure the goodness of clustering.

The Silhouette index ranges between -1 and 1, where high values indicate a well-matched point to its own cluster and poorly matched to neighboring clusters. If many points have negative values, the number of clusters should be adjusted. The optimal number of clusters is determined using the Silhouette Elbow method, which identifies the point at which the rate of improvement in the Silhouette index slows down.

However, the K-means method is inherently linear [60] and unsuitable for complex non-linear data distributions. To account for data non-linearity, we use a deep convolutional neural network (DCNN) for feature extraction. The DCNN transforms input heatmaps into feature vectors, which are more easily separable by a linear clustering algorithm [61] than the original heatmap.

The state-of-the-art InceptionV3 based on transfer learning is employed as the DCNN in this research. The InceptionV3 architecture is specifically designed to improve adaptability to different scales and prevent overfitting [62]. Transfer learning allows us to transfer pre-trained model parameters to our new model, thereby accelerating its training [63]. Utilizing a DCNN trained on a large dataset, such as ImageNet, enables the extraction of generic features applicable to other

images, such as heatmaps, without the need for training from scratch. Furthermore, the pre-trained DCNN weights enhance the accuracy of specific tasks, such as pattern recognition, when the available training data is limited [64]. In this study, we extract feature vectors from the demand heatmaps using the InceptionV3 deep neural network pre-trained on the ImageNet dataset, which contains millions of images for object recognition and image classification [65]. The compatibility of K-means with the linearly separable data produced by this feature extraction method, further justifies its use, as it balances the analytical approach by complementing the complexity of DCNN.

We use the Inception V3 architecture, pre-trained on ImageNet, to extract latent feature vectors from these residual heatmaps. By passing the residual matrix through the Inception V3 network, we obtain a latent representation—a vector—that preserves the essential information contained in the original matrix form. While this latent vector is not fully interpretable in terms of direct physical meaning, it provides a lower-dimensional feature space in which the residual patterns are more easily separable. This transformation allows us to apply K-means clustering to identify dominant temporal residual patterns more effectively than if we clustered the raw matrices directly.

Ultimately, this step transforms residual heat maps into feature vectors clustered based on temporal similarity. This process yields K clusters with distinct temporal residual patterns and determines the number of such clusters or patterns.

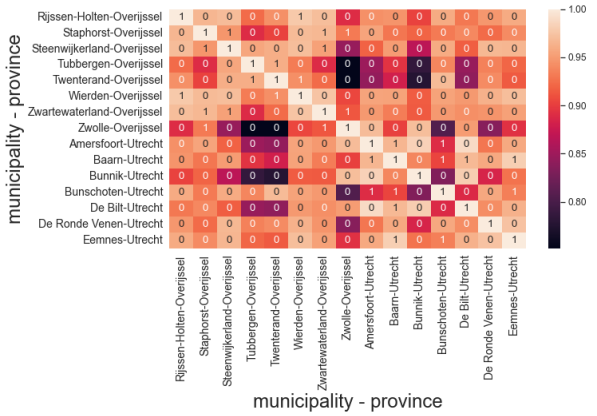
F. Association of Residual Patterns With Socio-Spatial Variables

In the final step, we analyze the degree to which the K clusters per spatial scale are associated with demographic and land-use variables. We assess the extent to which the socio-spatial characteristics of a TAZ can predict the cluster of temporal residual patterns it belongs to. We compare the distribution of various points of interest (POI) and demographics (derived from CBS) within the clusters to relate the resulting clusters (i.e., temporal residual patterns) to land use and socio-economic characteristics. Assuming a non-linear relationship, we propose a tree-based ensemble machine learning method, eXtreme Gradient Boosting (XGBoost) Ensemble, to model the relationship between land-use features and clusters. The regularization term in the loss function allows for achieving the lowest complexity with the highest accuracy.

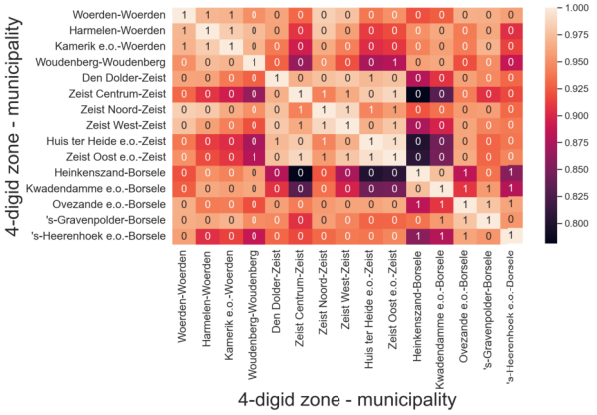
To identify the most important socio-spatial features contributing to residual patterns (i.e., clusters), we use the "gain" metric. In XGBoost, the gain metric represents the improvement in accuracy brought by a feature to the branches it is on. Essentially, it measures the importance of a feature in the model by calculating how much the feature contributes to the overall performance of the model. For more details on the XGBoost algorithm used in this study, we refer readers to Appendices A and B and [66].

III. Results and Discussion

In this section, we first present an initial exploration of trip production data and the predictions generated by the LSTM and LSTM+GCN models. In the second part of this section, we analyze the prediction results of the LSTM+GCN model in more detail. Here, we aim to identify residual patterns at each spatial scale and associate them with the socio-spatial features of each TAZ within the respective spatial scale.



(a) Correlation matrix between municipalities.

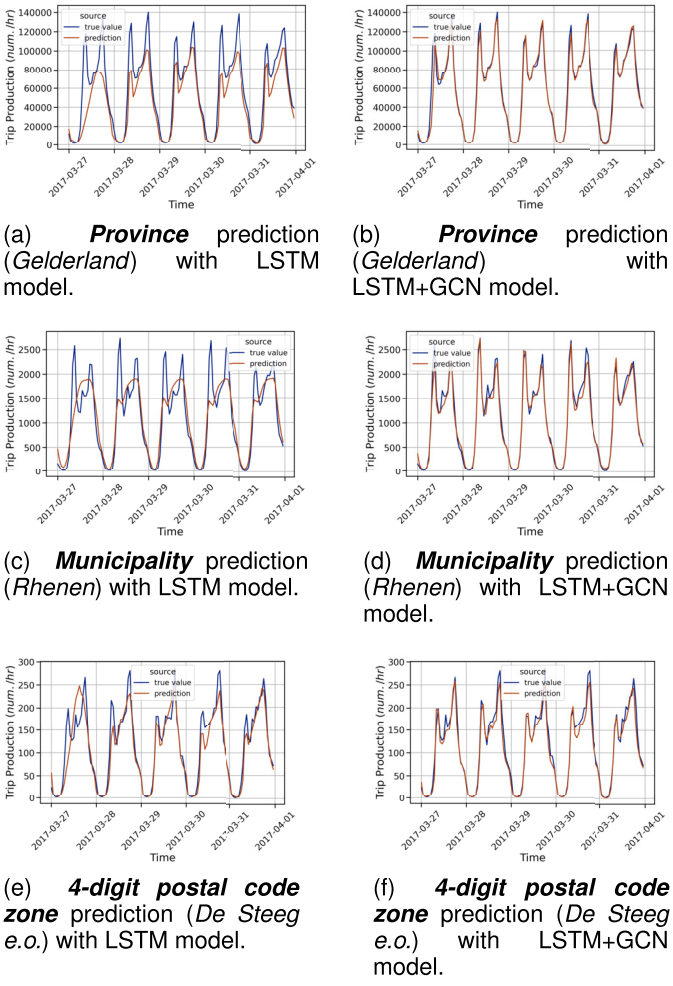


(b) Correlation matrix between 4-digit postal code zones.

Fig. 6. Correlation of trip production between TAZs under different spatial scales.

A. Initial Exploration and Predictions: LSTM Vs. LSTM+GCN Models

In order to investigate whether adjacent zones have similar trip production patterns, a correlation matrix is presented in Figure 6. Specifically, the matrix examines the correlation between several municipalities (Figure 6a) and several 4-digit postal code zones (Figure 6b). Each cell in the matrix contains a value of either 1 or 0, indicating whether the two associated zones are adjacent or not. Despite the observation that lower correlated TAZs are primarily non-adjacent, the data indicates no significant linear correlation between adjacent zones, as evidenced by the accompanying figures. However, a conditional non-linear correlation may exist between adjacent TAZs, which the linear correlation matrix fails to capture. To overcome this limitation, non-linear models, such as deep neural networks, are employed in this study to model complex behaviors that may exist between adjacent zones.



(a) ***Province*** prediction (*Gelderland*) with LSTM model.

(b) ***Province*** prediction (*Gelderland*) with LSTM+GCN model.

(c) ***Municipality*** prediction (*Rhenen*) with LSTM model.

(d) ***Municipality*** prediction (*Rhenen*) with LSTM+GCN model.

(e) ***4-digit postal code zone*** prediction (*De Steeg e.o.*) with LSTM model.

(f) ***4-digit postal code zone*** prediction (*De Steeg e.o.*) with LSTM+GCN model.

Fig. 7. The worst predictions of trip production based on the lowest MAPEs in the network among the three spatial scales using LSTM (on the left side) and their associated prediction using LSTM+GCN (on the right side).

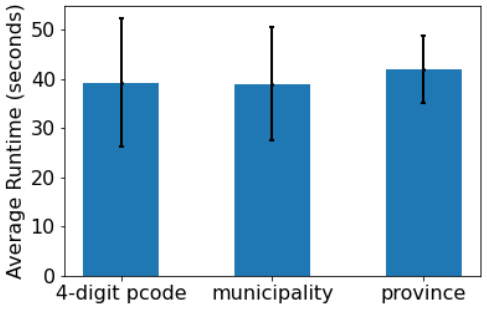


Fig. 8. Run time per TAZ of the LSTM model for the three spatial scales.

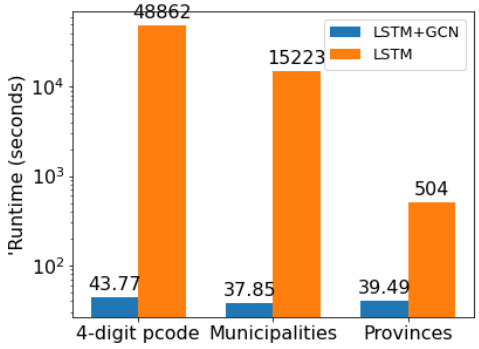


Fig. 9. Total run time of the LSTM and LSTM+GCN model for the three spatial scales.

In our modeling process, the first step was to pre-process the data to conform to the required format for further analysis. This step acquired a normalized trip production vector for each TAZ within one of the three studied spatial scales. We then applied the LSTM and LSTM+GCN models and evaluated the LSTM model's predictions based on its MAPE evaluation metric. Figure 7 presents the worst TAZs in the three spatial scales under study, along with the LSTM+GCN model's predictions for those TAZs. Including adjacency data in the LSTM+GCN model seems to improve the accuracy of predictions, particularly for TAZs where the LSTM model struggled to capture peak trip production values. LSTM seems to produce higher errors in predicting regular peaks, especially at the *municipality* spatial scale, and might need larger training/validation dataset. LSTM+GCN, on the other hand, seems to perform more accurately in predicting the daily peaks with the same amount of training data.

Although we observe improvements in the LSTM+GCN model's accuracy, the latent factors contributing to these results are not immediately apparent. Further research into these models and the properties of each TAZ in the network is necessary to gain insights into these factors. To comprehensively evaluate the results of our two models, we examined the distribution of the two evaluation metrics across the understudy scales, as shown in Figures 10 and 11. These figures indicate that the LSTM+GCN model's average MAPE and MSE of predictions are lower than those of the LSTM model across all three scales under study.

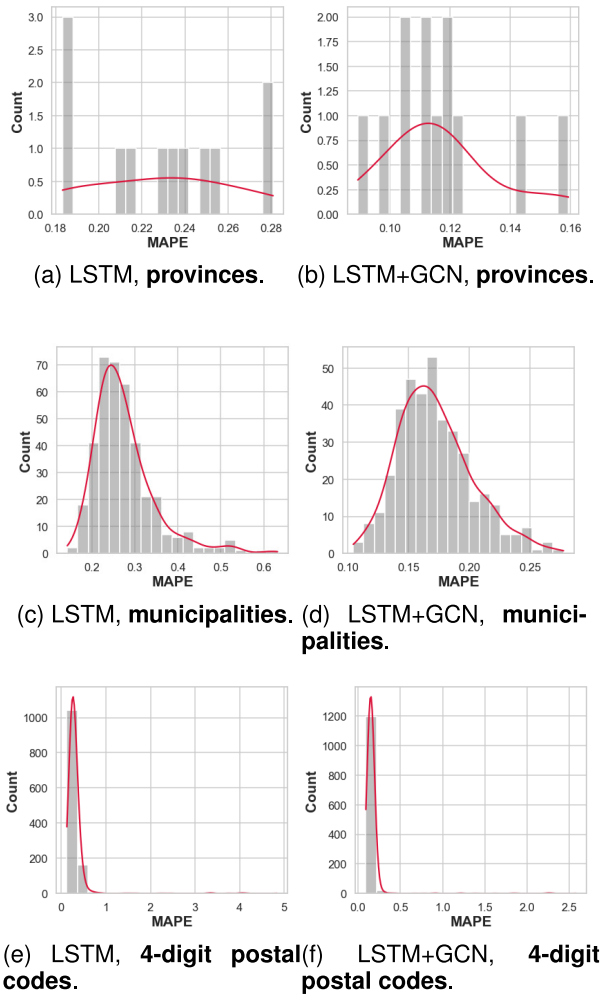
Table II presents the summary of prediction metrics of MAPE and MSE across different spatial scales using LSTM and LSTM+GCN model. The empirical results presented in this table engender a thought-provoking dialogue about the expected trends in predictive performance across varying spatial resolutions. In an intuitive sense, the forecast accuracy is anticipated to deteriorate with the increase in spatial granularity due to the amplified noise and intrinsic variability within the travel patterns. However, the LSTM model's performance metrics indicate an anomalously higher error for the municipality scale compared to the 4-digit postal code zones. This inversion of expected error magnitude necessitates a closer examination of underlying dynamics.

One plausible explanation for this counterintuitive finding could be the heterogeneity of socio-spatial characteristics within municipalities. The larger standard deviation in MAPE and MSE at the municipality level suggests a wider variability in prediction performance, which could stem from diverse commuting behaviors, land-use configurations, and transportation network complexities within these larger regions. Municipalities often encapsulate a mix of urban, suburban, and possibly rural settings, each with distinct travel demand

TABLE II

SUMMARY OF PREDICTION METRICS ACROSS DIFFERENT SPATIAL SCALES USING LSTM AND LSTM+GCN MODELS, WITH MAPE EXPRESSED IN PERCENTAGE UNITS. THIS TABLE CONTRASTS OUR BASELINE LSTM WITH THE PROPOSED LSTM+GCN. ACROSS ALL THREE SPATIAL RESOLUTIONS THE HYBRID MODEL REDUCES MAPE BY 47–99 % AND CUTS MSE BY 38–74 %, CONFIRMING ITS SUPERIOR PREDICTIVE POWER

Model	Metric	Province (Avg ± Std)	Municipality (Avg ± Std)	4-digit Zones (Avg ± Std)
LSTM	MAPE	23.0% ± 3.4%	2.6e+14% ± 5.2e+15%	1.3e+14% ± 3.5e+15%
	MSE	2.4e+08 ± 3.6e+08	5.5e+05 ± 3.5e+06	5.4e+04 ± 3.7e+05
LSTM+GCN	MAPE	12.0% ± 1.9%	8.7e+11% ± 1.7e+13%	1.6e+13% ± 5.6e+14%
	MSE	5.0e+07 ± 6.7e+07	1.4e+05 ± 1.4e+06	1.5e+04 ± 9.2e+04

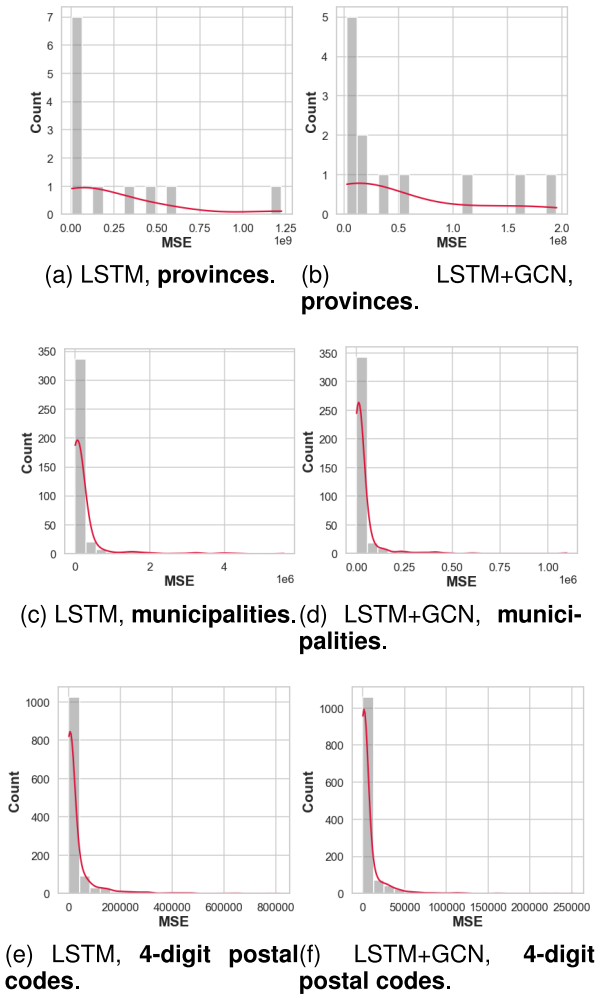


(a) LSTM, **provinces.** (b) LSTM+GCN, **provinces.**

(c) LSTM, **municipalities.** (d) LSTM+GCN, **municipalities.**

(e) LSTM, **4-digit postal codes.** (f) LSTM+GCN, **4-digit postal codes.**

Fig. 10. MAPE for actual versus predicted trip production values per TAZ for each of the three spatial scales, using LSTM (left) and LSTM+GCN (right) model.



(a) LSTM, **provinces.** (b) LSTM+GCN, **provinces.**

(c) LSTM, **municipalities.** (d) LSTM+GCN, **municipalities.**

(e) LSTM, **4-digit postal codes.** (f) LSTM+GCN, **4-digit postal codes.**

Fig. 11. MSE for actual versus predicted trip production values per TAZ for each of the three spatial scales, using LSTM (left) and LSTM+GCN (right) model.

patterns that could potentially confound the LSTM model, which does not explicitly account for spatial adjacency.

Conversely, the LSTM+GCN model, which integrates spatial adjacency into its predictive framework, exhibits a substantial improvement in MAPE for municipalities, although the MSE remains higher than that for the 4-digit zones. The improved MAPE implies that when spatial relationships are considered, the model becomes more adept at capturing the proportionate variances in trip production, especially in the context of municipalities where spatial adjacency plays a significant role. The persistence of higher MSE, despite a lower MAPE, might suggest that while the model is generally accurate, it is occasionally susceptible to larger errors—potentially from extreme values or outliers that are more pronounced in municipal data.

These results underscore the merit of integrating spatial adjacency to enhance predictive accuracy, especially where the spatial structure itself may be a pivotal determinant of travel patterns. The LSTM+GCN model's ability to leverage such spatial correlations ostensibly attenuates the prediction difficulty at higher spatial scales. Nonetheless, the nuanced nature of trip production across different spatial scales reaf-

firms the necessity for tailored-modeling approaches that can accommodate the unique features of each granularity level.

As we move from the *province* scale to lower levels of abstraction, the prediction accuracy declines, as indicated by increasing MAPE values. This trend suggests that prediction accuracy is lower at lower levels of abstraction. Nonetheless, it appears that accuracy improves at higher levels of abstraction, despite the increase in spatial uncertainty. This outcome may be because the aggregation of trip production patterns makes them more regular and hence more predictable. The worst predictions for both models are presented in Figure 7 and support our observation that the LSTM+GCN model predicts peak trip production values more accurately than the LSTM model. Furthermore, the worst predictions demonstrate that trip production at the *province* scale is more accurately predicted than at higher resolution scales. This outcome may be due to pattern aggregation and resolution lowering.

Our study's results indicate that incorporating adjacency data into the LSTM+GCN model enhances the accuracy of extreme value predictions while also significantly decreasing computation time. Figure 8 shows that the computation time distribution for the LSTM model has an average of 40 seconds per TAZ, with a more extensive interquartile range for 4-digit postal code zones, suggesting a more diverse range of zones. The total computation times for the LSTM and LSTM+GCN models are illustrated in Figure 9. We noticed that the LSTM model's run time depends on the number of TAZs in the network, whereas the LSTM+GCN model's computation time does not fluctuate significantly with the number of TAZs. This is because the LSTM+GCN model is only trained once, rather than training one model for each TAZ in each spatial resolution. Therefore, changes in run time under different spatial scales are not substantial.

Considering an alternative approach, one could estimate a single LSTM model for all TAZs without incorporating adjacency data or using the GCN model. This method would reduce computation time since it would no longer scale with the number of TAZs. However, the impact on prediction accuracy might differ between the disaggregate LSTM and aggregate LSTM+GCN models. Lacking the spatial relationships between TAZs, the single LSTM model might experience a decrease in prediction accuracy, as it would not capture the spatial heterogeneity and correlations among TAZs that influence trip production patterns. Regarding prediction accuracy, the single LSTM model without adjacency data would likely fall between the disaggregate LSTM models, which can capture the unique characteristics of each TAZ, and the aggregate LSTM+GCN model, which benefits from the inclusion of adjacency data to capture spatial relationships better. Thus, while using a single LSTM model for all TAZs without incorporating adjacency data offers the advantage of reduced computation time, it may involve a trade-off in prediction accuracy compared to the other approaches.

Altogether, our findings imply that incorporating adjacency data can improve prediction accuracy while decreasing computation time. Consequently, we analyzed and explored the predictions made using LSTM+GCN for the remainder of our study.

It is important to note that our empirical observations indicate that using a single model trained on the entire network may be more convenient than training separate models for each TAZ. However, we have not performed a formal computational complexity analysis or extensive runtime evaluations, and thus any conclusions about computation time savings remain preliminary.

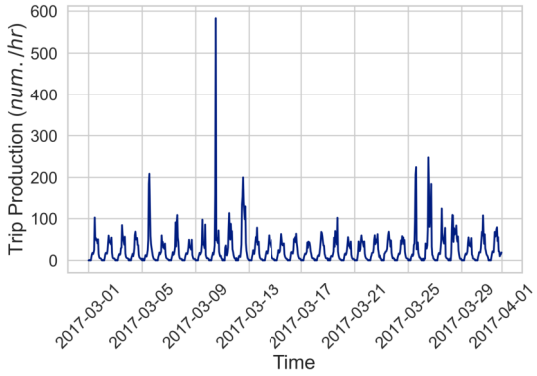
Figures 10 and 11 visualise the dispersion of prediction error over the demand spectrum. Figure 10 corresponds to the *MSE*; Figure 11 to the *MAPE*. These systematic patterns motivate the residual-clustering exercise of the next part, in which we further examine whether such outliers share common socio-spatial attributes.

B. In-Depth Analysis of LSTM+GCN Prediction Results: Residual Patterns and Socio-Spatial Features

In Figure 12, we present the trip production and prediction results for Vlieland, a northern island that is both a municipality and a 4-digit zone. It is worth noting that Vlieland holds the highest MAPE among all 4-digit zones and municipalities, indicating a considerable challenge in accurately forecasting trip production in this region.

The observed trip production pattern of Vlieland in Figure 12a is characterized by an irregular profile with an extreme peak in the first half of the month, followed by scattered high production in the last days of the month. This erratic pattern might be attributed to special events associated with this TAZ. Figure 12b and 12c present the trip production and prediction at the municipality and 4-digit zone levels, respectively. Accordingly, the prediction performance of the models varies across different spatial scales. Although the two models share similar hyper-parameters, they are trained on different training sets due to aggregation at a higher level of abstraction. The municipality-level model seems to capture the time series with more variation in the values, implying better predictions of production values influenced by seasonality. Conversely, the 4-digit model appears to be more biased towards predicting values closer to the average production line, hence less sensitive to extreme values in the production but better at predicting the off-peak values. These observations suggest the importance of considering the spatial scale in trip production and prediction analysis, as it can significantly affect the accuracy of the results.

Figure 13 presents the average hourly trip production and the prediction residual for various provinces in the Netherlands. This figure depicts these two variables, with Figure 13a showing the average hourly trip production and Figure 13b displaying the average hourly trip production prediction residual. The figure highlights that provinces such as Noord-Holland, Zuid-Holland, and Noord-Brabant, characterized by high population density and heavy industrial and commercial areas, have higher trip production values with high variability. This variability is shown in the shaded areas between the 10th and 90th percentile of the values. Figure 13b implies that residual peaks occur around the peak hours of trip production, although with somewhat different patterns. For instance, Zuid-Holland has the highest average hourly trip production during the afternoon rush hour, while the residual peak happens

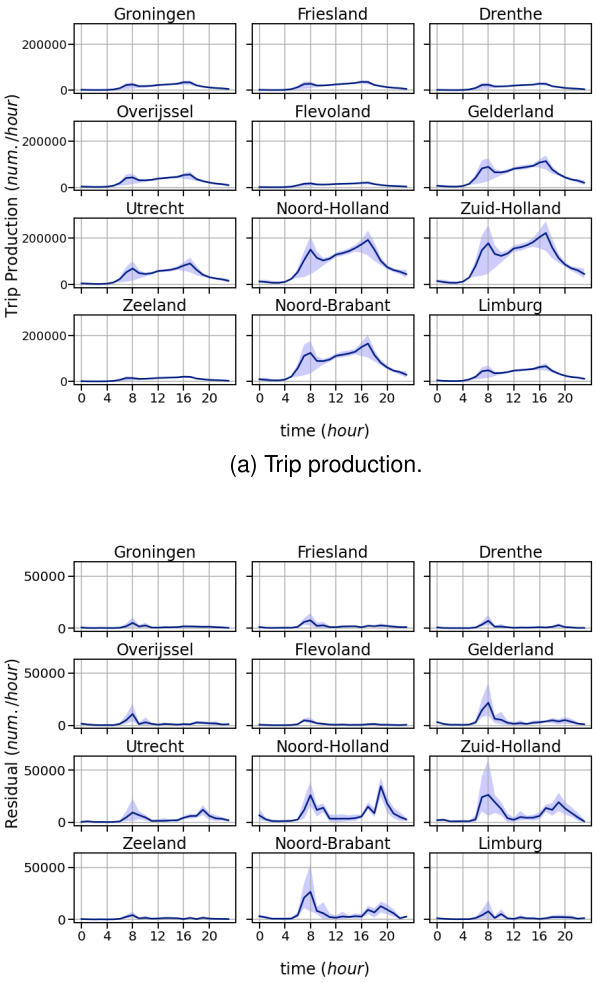


(a) Actual trip production of Vlieland throughout March 2017.

(b) Prediction of trip production for Vlieland at the **municipality** level.

(c) Prediction of trip production for Vlieland at the **4-digit postal code** level.

Fig. 12. Vlieland trip production prediction under two spatial scales throughout the last working week of March 2017.



(a) Trip production.

(b) Prediction residual.

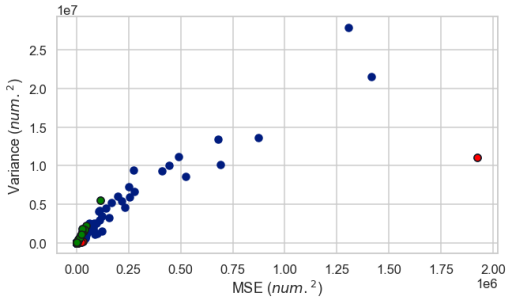
Fig. 13. Average hourly trip production and prediction residual time series among the provinces in The Netherlands.

during the morning peak, which has the highest variation in trip production. These observations suggest that the prediction error is correlated with the variation in trip production, which holds implications for improving the accuracy of transportation demand models. However, further research is needed to explore these correlations and their underlying causes in more detail.

To investigate the correlation between trip production prediction error and variation in trip production among the 4-digit zones, Figure 14a plots the Mean Squared Error (MSE) against variance. The figure reveals a reasonably linear correlation between the two variables, indicating that the prediction error is positively correlated with trip production variation. However, a few outliers with high MSE and relatively low variance suggest high prediction errors despite low variation in trip production. Figure 14 highlights these TAZs in red, pointing to the zones with a high prediction error despite low variation in trip production. Interestingly, all the islands are among the outliers.

The robustness of the proposed LSTM+GCN model is crucial given the diversity of urban and rural dynamics in different regions. Our model was trained and evaluated on regions with varied geographical and traffic conditions, specifically at multiple spatial scales including provinces, municipalities, and 4-digit zones. The model consistently demonstrated its ability to capture spatial-temporal dynamics effectively, achieving high predictive accuracy across urbanized and rural regions.

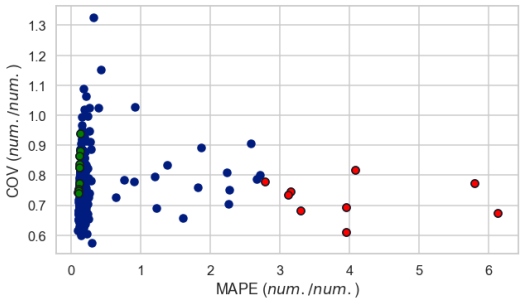
Despite the expected variations in traffic patterns between densely populated urban regions and sparsely populated rural regions, the model achieved consistent predictive accuracy in all scenarios. This was particularly notable during rush hour



(a) MSE of trip production prediction vs. the variance of trip production.

(b) Zones with the highest MSE/variance ratio.

Fig. 14. MSE of trip production prediction vs. the variance of trip production among 4-digit zones in The Netherlands.



(a) MAPE of trip production prediction vs. the COV of trip production.

(b) Zones with the highest MAPE/COV ratio.

Fig. 15. MAPE of trip production prediction vs. the COV of trip production among 4-digit zones in The Netherlands.

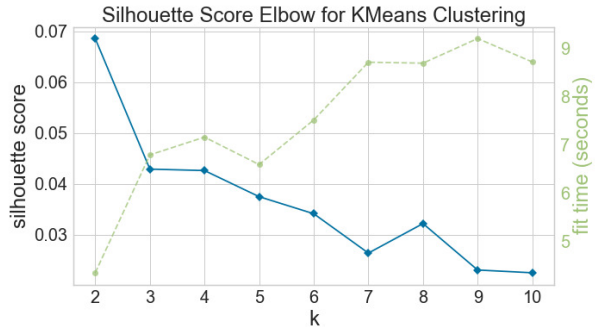
periods and off-peak periods, as well as across weekday and weekend traffic. Leveraging the complementary strengths of both GCN and LSTM, the model could effectively capture spatial and temporal correlations even in regions with varied urban dynamics.

These results highlight the robustness and generalizability of the LSTM+GCN model, making it suitable for trip production prediction across diverse geographical areas and traffic conditions.

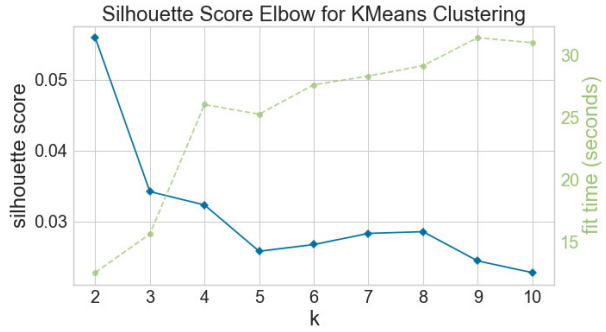
To investigate the relative prediction error, Figure 15a explores the relationship between Mean Absolute Percentage Error (MAPE) and Coefficient of Variation (COV) in trip production among the 4-digit zones. The figure indicates a positive correlation between MAPE and COV, suggesting that the prediction error increases with an increase in trip production variation. However, some outliers with high MAPE and low COV suggest high prediction errors despite low variation in trip production. Figure 15 highlights these TAZs in red. Overall, the results suggest that the prediction of trip production becomes more challenging in certain zones due to underlying characteristics.

In this section, we investigate the patterns and characteristics of trip production residuals. To accomplish this, we generated a heatmap of trip production residuals for each TAZ. The heatmap displays the normalized prediction residual of the associated TAZ for each hour of the day (y-axis) over the working days of the week (x-axis) as displayed in Figure 5 in the previous section. Each TAZ has its specific day-to-day and within-day residual patterns. To better understand the dominant patterns of residuals, we analyzed the heatmaps of each spatial level separately, i.e., municipality and 4-digit zones. We employed a DCNN for feature extraction, followed by the K-means clustering algorithm to identify the dominant patterns in each spatial scale. We used the Silhouette score elbow method to determine the optimal number of clusters. Figures 16a and 16b depict the Silhouette plots at the municipality and 4-digit zone levels, respectively. Based on these plots, we found four and five clusters suitable for the municipality and 4-digit zone levels, respectively.

Figure 17 shows the spatial distribution of identified clusters for TAZs in the Netherlands at different spatial scales: Municipality (17a) and 4-digit zone level (17b). At the municipality scale, clusters 0 and 1 appear to be concentrated on the west side of the Netherlands, while clusters 2 and 3 are more prevalent on the eastern side of the country. However, at the 4-digit level, a distinct pattern is not observable by solely examining the spatial distribution. These findings suggest that additional underlying factors might contribute to the

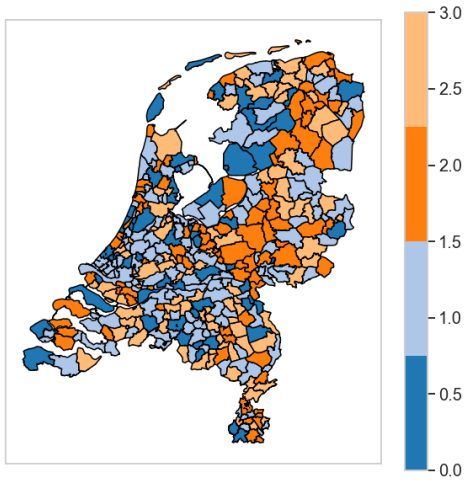


(a) Silhouette plot for municipalities.

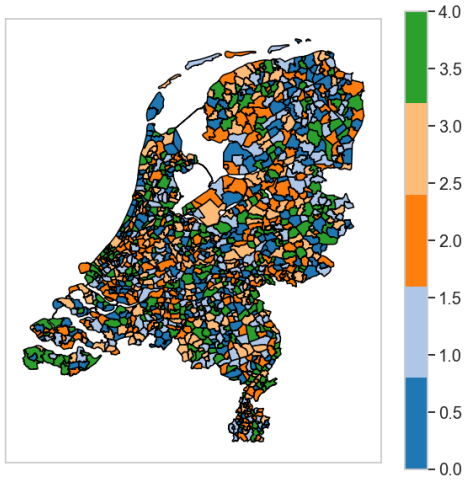


(b) Silhouette plot for 4-digit zones.

Fig. 16. Silhouette score elbow plots.



(a) Municipalities.



(b) 4-digit postal code zones.

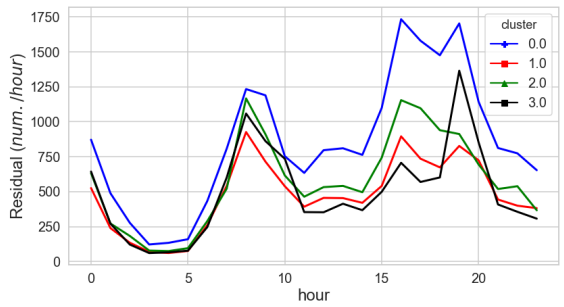
Fig. 17. Spatial distribution of clusters for TAZs in the Netherlands.

formation of production residual clusters and thus warrant further analysis.

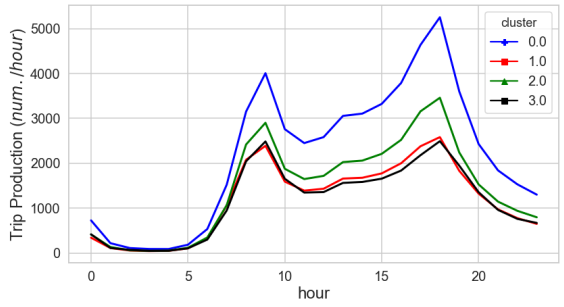
Figures 18 and 19 compare the residual patterns with the trip production patterns across identified clusters under the municipality and 4-digit spatial scales, respectively. This analysis aims to gain insight into the relationship between trip production and its associated prediction residual. The results shown in Figure 18 indicate that the most severe within-day trip production peak occurs commonly between all four clusters at around 6 p.m. or with a slight difference at around 9 a.m. However, the peak of residual is different among clusters. Specifically, cluster 0 displays the highest residual among all clusters, with the residual peak occurring at 4 p.m. and with a slight difference at 7 p.m., while the morning peak occurs at 8 a.m. In cluster 1, residual peak is at 8 a.m., and the afternoon peak occurs at 4 p.m. For cluster 2, the morning and afternoon peaks of the residual are equally severe, and they happen around 8 a.m. and 4 p.m., respectively. In cluster 3, the prediction residual peak is at 7 p.m., and the morning peak occurs at 8 a.m.

Overall, the identified clusters seem to relate to the districts where their prediction residual occurs and the severity of the morning and afternoon peaks. It is interesting to note that, unlike the trip production pattern, whose only afternoon peak occurs at 6 p.m., the residual patterns display a double afternoon peak at 4 and 7 p.m., i.e., 2 hours before and one hour after the production peak in the afternoon. However, the morning peak of residual happens at the same time, around 8 a.m., meaning one hour before the production morning peak. Cluster 0, with the highest residual average at all hours of a day, has a double severe afternoon peak, suggesting that high production results in more irregularities scattered before (i.e., from two hours before) and after (i.e., to one hour after) the actual afternoon peak of production. Moreover, morning activities seem to follow a more strict schedule, as the residual elevation is scattered in a narrow range (i.e., from one hour before to the peak hour of trip production) among all four clusters at this scale. Except for the peak values of residual, the residual patterns follow the discerned patterns in the trip production.

Figure 19 illustrates the average hourly residual clusters at the 4-digit level and compares it with the corresponding trip production clusters. The analysis reveals that the residual

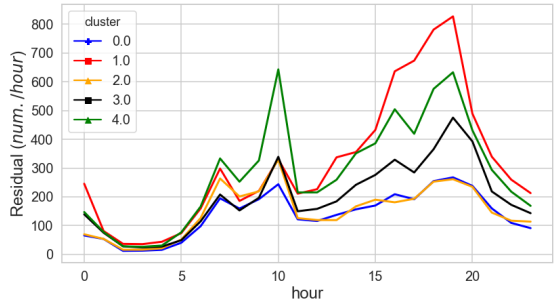


(a) Average hourly trip production residuals.

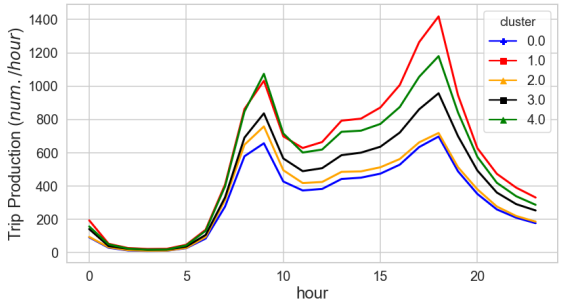


(b) Average hourly trip production.

Fig. 18. Comparing average hourly trip production and residuals across identified clusters at the **Municipality** Scale.



(a) Average hourly trip production residuals.



(b) Average hourly trip production.

Fig. 19. Comparing average hourly trip production and residuals across identified clusters at the **4-digit postal code** scale.

patterns differ from those identified at the municipality scale, indicating the importance of analyzing them to gain insight into the relationship between trip production and its associated prediction residual at different spatial scales. Specifically, the study finds that the patterns of trip production across these clusters are similar to those identified at the municipality scale, with the morning and afternoon peak of trip production occurring at around 9 a.m. and 6 p.m., respectively, across all clusters. However, unlike the municipality scale, the residual patterns at the 4-digit level exhibit double morning peaks occurring at 7 a.m. and 10 a.m., two hours before and one hour after the morning peak of trip production. Additionally, the afternoon residual peak is reduced to a single peak at around 7 p.m., one hour after the afternoon trip production peak. Further analysis of the five clusters at the 4-digit TAZs level reveals that the severity of the first and second morning and afternoon peaks differs between the clusters, with cluster 1 exhibiting the highest afternoon peak and cluster 4 having the most severe morning peak.

In our analysis, we employed the XGboost algorithm and used the "gain" metric to assess the importance of demographic and land-use features associated with the identified prediction residual clusters. "Gain" is a measure of the relative contribution of each feature to the model calculated as the average gain of the feature when it is used in all possible splits of a tree in the ensemble. The "gain" metric provides an intuitive measure of feature importance and identifies the most relevant features associated with the prediction residual clusters. The higher the gain value of a feature, the more important it is in predicting the outcome variable and the more significant its contribution to the model identifying its residual pattern.

This section presents the results of an importance analysis conducted using the "gain" metric in the XGboost algorithm to identify the most relevant demographic and land-use features that distinguish residual patterns in two spatial scales, municipality, and 4-digit. Box plots of the feature importance scores are displayed in Figures 20 and 21, where the y-axis represents the feature value, and the x-axis shows the cluster names. It should be noted that the plot displays the distribution of feature values among the TAZs within each cluster rather than representing the gain value itself.

Figure 20 shows the results for the municipality scale, where the identified features are mainly related to points of interest (POIs), including the average number of nearby primary schools, cafes, restaurants, and general practices, which tend to increase the residual and make trip production less predictable. In addition, the concentration of older or newer-built houses in the TAZ affects the prediction error, while the proportion of younger residents appears to make trip production more regular, as observed in cluster 0. Notably, cluster 3 exhibits a severe late afternoon peak in residual, an hour later than the trip production peak, with a higher average electricity consumption.

Figure 21 displays the importance analysis results for the 4-digit scale, where socio-economic features become more critical in distinguishing the residual cluster of TAZs. The first 12 most important features include residents with a non-EU immigrant parent, the percentage of high-income residents, residents with unemployment, social or disability benefits,

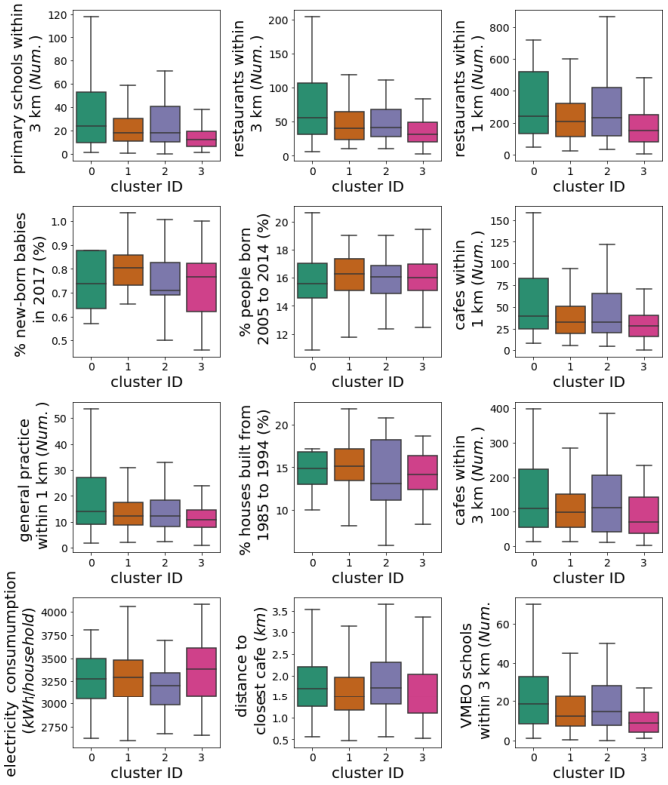


Fig. 20. Box plots of the most relevant features to the identified residual clusters at the municipality spatial scale.

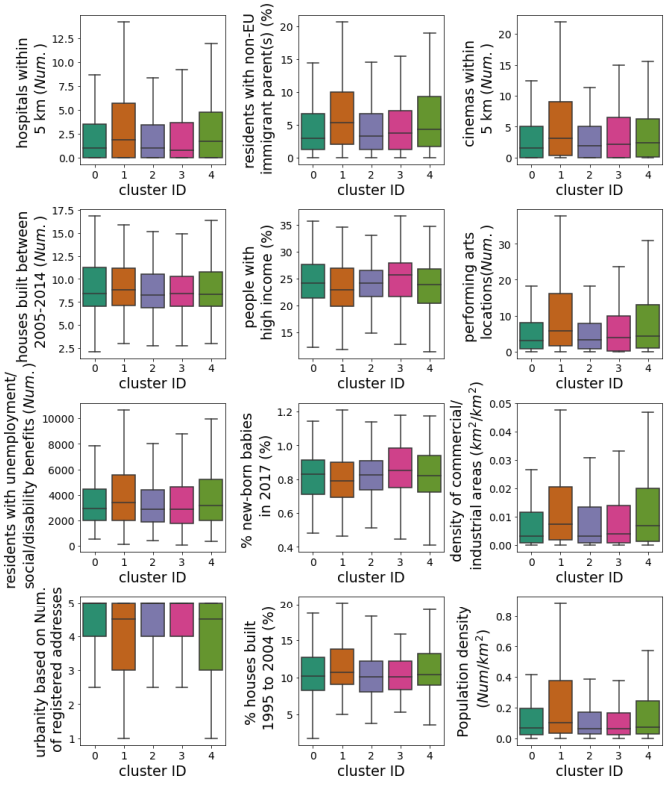


Fig. 21. Box plots of the most relevant features to the identified residual clusters at the 4-digit spatial scale.

the percentage of newborn babies, and population density. Land use and urbanity features also play an important role

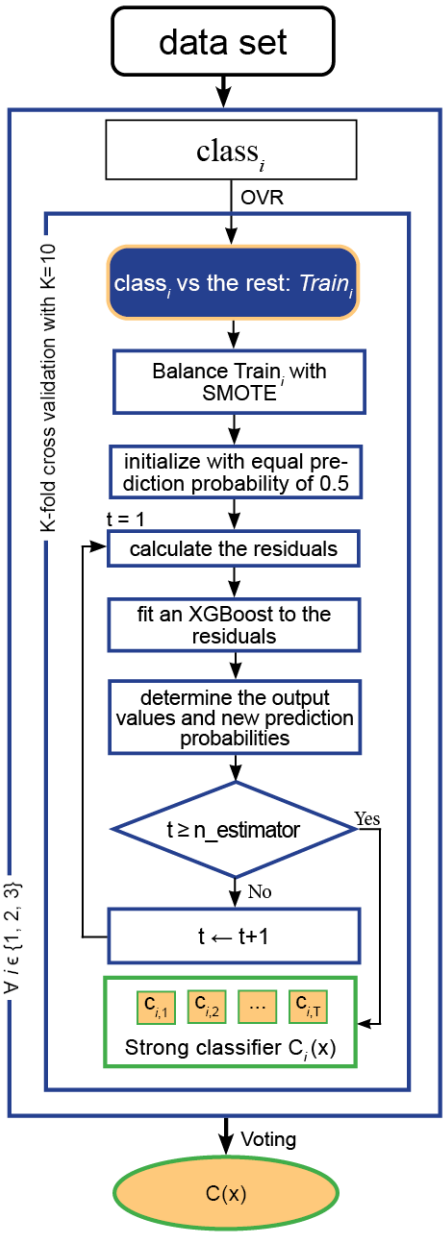


Fig. 22. The framework of the OVR-SMOTE-XGBoost ensemble model.

in distinguishing the residual cluster. For instance, the number of close hospitals, cinemas, and performing art locations, the density of commercial/industrial areas, the urbanity score, and the concentration of middle-aged houses are all relevant. Cluster 1, with the highest afternoon peak in residual, seems to have more close hospitals, a higher proportion of residents with a non-EU immigrant parent, and entertainment locations such as cinemas and performing art venues, as well as a higher percentage of residents with social benefits.

Overall, the results suggest that the prediction of trip production relies more on demographic information when analyzing at lower levels of abstraction, whereas, at higher levels of abstraction, spatial features such as land use and built environment variables play a more critical role in causing irregular demand patterns.

C. Research Limitations

This research, while contributing significant insights into the field of short-term trip production prediction, acknowledges certain limitations that should be considered when interpreting the results and applying the methodologies.

1) The primary data source for this study is derived from GSM traces provided by a major Dutch telecommunications company, Vodafone. While these traces offer valuable insights into travel patterns, they inherently come with significant limitations that might affect the study's outcomes. The data covers approximately one-third of the Dutch population, primarily Vodafone users. This coverage limitation potentially introduces biases, as the user demographic may not accurately represent the wider population's travel behavior, especially across different socioeconomic groups. Also, the method used to infer trip productions from GSM data relies on assumptions about travel modes, which may not always hold true. The accuracy of these inferences can be affected by the granularity of the data and the algorithms used to interpret signal patterns, possibly leading to misclassifications (e.g., distinguishing between travel by public transport and private vehicles). Furthermore, the accuracy of GSM-based studies is highly dependent on the spatial distribution and density of network cells. Areas with sparse cellular coverage or where cell towers are unevenly distributed can lead to significant gaps in data, affecting the reliability of trip estimations. Additionally, the data's temporal coverage (e.g., time of day, day of the week) can also influence accuracy. For instance, trip productions during peak traffic hours might be overrepresented due to higher mobile phone usage, skewing the understanding of typical travel patterns. Given these constraints, particularly the black-box nature of the processed GSM data, this study acknowledges the limitations in the accuracy and reliability of the trip production values derived. While the Dutch case offers a complete, nationwide test bed, we recognize that its specific spatial structure could limit external validity. We therefore explicitly invite future work to benchmark the proposed framework on widely used public corpora in order to assess robustness across heterogeneous geographies.
2) The scope of this study is geographically limited to the Netherlands. While this provides detailed insights into Dutch urban and transportation network, the findings might not directly translate to other regions with different geographic, demographic, and infrastructural characteristics.
3) The study adopts LSTM integrated with GCN as the primary predictive model. This choice, while based on the model's ability to handle spatial-temporal correlations and variations, does not imply it is the optimum method for all trip production prediction scenarios. The field offers a variety of other data-driven models that might provide different levels of effectiveness depending on specific use cases.
4) Another limitation of our study is the lack of comparison with a broader range of state-of-the-art models. While our focus was on demonstrating the benefits of integrating spatial adjacency into LSTM models, future research should include comparisons not only with models like T-GCN [42] and STGCN [36] but also with advanced graph-based architectures that incorporate dynamic graphs [43], transformer-based GCNs [44], and hierarchical GCNs (HGCNs) [45]. Such evaluations would provide a more complete perspective on how effectively our approach captures spatial dependencies relative to more complex models.
5) The decision to use K-means clustering for residual analysis was guided by its simplicity and effectiveness. However, this approach might not capture the intricacies of more complex, non-linear data distributions compared to other advanced clustering algorithms.
6) The LSTM+GCN model, though efficient in handling large-scale data, involves a certain level of computational complexity. This aspect might limit its applicability in environments with constrained computational resources.
7) The findings and conclusions drawn from this study, particularly those related to the residual analysis and socio-spatial feature associations, are based on the specific context of the Netherlands. The transferability of these insights to other contexts requires a separate study.
8) A limitation of our study is the absence of more granular socioeconomic data, such as income levels, education levels, and job types, in our analysis. Incorporating these detailed socioeconomic variables could potentially enhance the model's predictive capabilities and provide deeper insights into trip generation patterns. Future research could explore the integration of such data to analyze their impact on trip generation forecasts.
9) Another area for future research is the interpretability of the GCN+LSTM model. Employing techniques such as attention mechanisms, feature importance analysis, or explainable AI methods could provide deeper insights into the model's decision-making process and identify key influencing factors, enhancing transparency and trust in the model's predictions.
10) While our model shows promise, practical deployment for real-time prediction adjustments involves challenges not addressed in this study. Future work could focus on optimizing the model for real-time applications, exploring techniques to improve computational efficiency, and evaluating its performance in operational environments. Addressing these considerations would enhance the model's practicality for policymakers and transportation planners.
11) While we discuss certain computational aspects of our approach, such as training a single model for all TAZs instead of one model per TAZ, our primary contribution lies in demonstrating the incremental predictive accuracy gained by integrating spatial adjacency. We did not conduct a formal theoretical computational complexity analysis or extensive runtime evaluations. Future

research could focus on detailed computational profiling, optimizing the method for real-time deployment, and comparing its computational performance against more computationally efficient models.

12) Our residual analysis utilized K-means clustering to identify patterns; however, more sophisticated methods such as hierarchical clustering, DBSCAN, or spatial econometric approaches might provide deeper insights into residual patterns. Future research could explore these methods to better understand the factors contributing to prediction inaccuracies and potentially improve upon the results of our existing analysis.

To sum it up, this study's primary contributions lie in the multi-scale analysis of trip production prediction and the exploration of residual patterns and socio-spatial features across different spatial scales. While acknowledging the aforementioned limitations, the research provides a foundational framework that can be built upon and adapted for further studies in this domain. Future research could address these limitations by incorporating a broader data scope, exploring alternative predictive models, and adapting the methodologies to diverse geographical and urban contexts.

IV. Conclusion

This study presented a GCN+LSTM framework that integrates spatial adjacency into trip production prediction across multiple spatial scales. We demonstrated incremental gains in predictive accuracy by considering spatial heterogeneity and identified socio-spatial features critical to understanding residual patterns. These findings support the development of more tailored, accurate, and practical demand prediction models.

The findings of this research have several implications for understanding how spatial heterogeneity affects demand prediction and for transportation planning and policy-making. The results of this study are helpful for transport modelers to consider the spatial scale in selecting the relevant types of variables used in their models for travel demand prediction. For instance, recognizing that at lower levels of spatial abstraction, a combination of land-use and demographic information contributes significantly to residual patterns, while at higher spatial scales, land use plays a more critical role. This awareness allows for more tailored and accurate modeling approaches, leading to better-informed decisions regarding resource allocation, infrastructure development, and service planning.

Incorporating spatial adjacency into trip production prediction models using GCN has demonstrated improvements in prediction accuracy and computational efficiency. This enhancement enables transportation planners to efficiently handle large-scale networks without compromising accuracy, facilitating the development of more effective transportation strategies. Understanding the influence of neighboring TAZs on trip production can help in designing more efficient transit networks and optimizing traffic management strategies that account for spatial interactions between regions.

Moreover, our analysis of prediction errors, residual patterns, and their association with socio-spatial features helps identify areas where certain demographic or land-use characteristics contribute to demand variability. Policymakers can leverage this information to implement targeted interventions, such as adjusting land-use policies or enhancing transportation services in areas where prediction errors are linked to specific socio-spatial factors, ultimately better meeting the travel demands of different regions.

Although we focused on the prediction of trip production, the insights gained from this study assist in OD matrix estimation and prediction. Furthermore, while our study focused on motor-vehicle OD matrices, addressing trip production patterns specific to motor vehicles, it is essential to note that the methodology employed in this research is mode-agnostic. This versatility implies that our approach could potentially be adapted and applied to various other modes of transportation, broadening the scope and impact of the findings presented in this paper.

Overall, this research lays the groundwork for developing more complex demand prediction models with higher accuracy without requiring high computational capacity. By refining models to consider nuanced socio-spatial dynamics, we enhance their predictive accuracy and applicability, supporting the creation of more responsive and equitable transportation systems.

The study also enhances our understanding of spatial uncertainty across multiple spatial scales and how it affects the prediction of trip production. However, a major limitation of this study is that the adjacency of the TAZs is the only feature used for generating the adjacency matrix in our GCN. Other socio-spatial features, as we showed in this research, also affect the heterogeneity of trip production in areas. For instance, the spatial urbanization level, defined based on the combined land-use characteristics and demographics of an area, has been shown to affect trip production patterns and cause heterogeneity. Therefore, there is a definite need to consider other contributing spatial features for defining the adjacency matrix and comprehensively defining a dynamic (showing that features can change over time) adjacency matrix to address trip production heterogeneity.

Compared to previous studies that utilized traditional time series models or standalone deep learning models, our GCN+LSTM model demonstrates improved prediction accuracy by effectively capturing spatial-temporal dependencies. For instance, studies like Zhao et al. [42] and Yu et al. [36] also highlighted the benefits of incorporating spatial information. Our findings align with these results and further extend them by analyzing multiple spatial scales and focusing on trip production prediction at a national level.

As a future direction, we are committed to exploring cutting-edge methodologies in the domain of demand prediction, with a keen focus on benchmarking an array of forecasting paradigms. This includes delving into models that intricately weave together more sophisticated spatial data representations alongside advanced machine learning algorithms. Our objective is not merely to augment the predictive precision of these models but to catalyze a transformative shift in their capability to anticipate travel demand dynamics accurately. By charting this course, we aspire to contribute to the perpetual enhance-

ment of forecasting models, ensuring they remain both relevant and robust in the face of evolving transportation landscapes and the complex interplay of spatial-temporal factors they encompass.

APPENDIX A
AOVR-SMOTE-XGBOOST ENSEMBLE MODEL

In this appendix, we describe the methodology of our OVR-SMOTE-XGBoost ensemble model, which is designed for multi-class imbalanced data. Figure 22 illustrates the framework of this algorithm, which consists of three main steps:

1) **Decompose the multi-class classification problem**: We transform the multi-class classification problem into multiple independent binary classification problems. Traditional classification methods are typically designed for binary problems, and the complexity of multi-class classification problems can be mitigated by breaking them down into several binary classification problems. Using the One-vs.-Rest (OVR) strategy, we develop multiple classifiers—one for each class—indicating whether or not an instance belongs to a specific class [67]. Based on the OVR decomposition method, we decompose the initial training set into three two-class training sub-samples: $Train_1$ for class1, $Train_2$ for class2, and $Train_3$ for class3. Each classifier calculates the probability of each class using a binary logistic loss function. Instances are then classified into the class with the highest probability.
2) **Balance the training sets**: Imbalanced learning is a common problem in classification tasks, where underrepresented data and class distribution skew can negatively impact algorithm performance [68]. To address this issue, we employ the Synthetic Minority Oversampling Technique (SMOTE), a well-known data augmentation method that balances class distribution by oversampling minority class instances through random replication [69].
3) **Train the SMOTE-XGBoost ensemble model**: For each of the three OVR classifiers, we use the balanced training sets to train the SMOTE-XGBoost ensemble model for binary class prediction of TAZs.

To avoid overfitting, we implement a K-fold cross-validation strategy with $k = 10$. K-fold cross-validation divides the dataset into K equally sized groups of samples, or folds, and iteratively trains the model on K-1 folds while testing on the remaining fold. For example, given a dataset of 100 samples, 90 samples (i.e., nine folds) are used for training and validation, while the remaining 10 samples (i.e., one fold) are used for testing. This process is repeated for all ten folds, with no overlap, to achieve a robust accuracy in prediction.

APPENDIX B
XGBOOST ENSEMBLE MODEL

In this section, we provide a detailed explanation of the XGBoost algorithm. For further information, please refer to [66]. XGBoost is an ensemble of decision trees that are sequentially developed, with each tree working to improve the performance of the previous tree [70]. Ensemble methods aim to reduce the bias or variance of several weak learners by combining them into a strong learner (i.e., a learner with low bias and variance). Boosting is an ensemble method where weak learners are fitted sequentially and aggregated to the ensemble model. In each step, the training set is updated to focus more on the weakness of the current ensemble. In other words, each model in the sequence does the fitting by giving higher weight to misclassified data points. If the weak learner of each step depends on the gradient direction of the loss function at each step, this method is also called Gradient Boosting Machines (GBM) [71]. The advantage of XGBoost over non-extreme gradient boosting methods is the regularization term in the loss function, which helps prevent overfitting. Suppose our dataset is $\chi = (x_i, y_i) : i = 1, \ldots, n, ; x_i \in \mathbb{R}^m, ; y_i \in \mathbb{R}$. We have n observations, each with m features corresponding to their associated label y. Then $\hat{y}_i$ can be defined as a result of an ensemble, with T additive functions, represented by the generalized model as follows:

$$\hat{y}i = \Phi(xi) = \sum^{T} t = 1ft(x_i); , \tag{4}$$

where f_t is a decision tree, and $f_t(x_i)$ is the score given by the $t-th$ decision tree to the $i-th$ data point. The objective function that needs to be minimized to select the function f_t consists of two terms: *training loss*, $L(y_i, \hat{y}i)$, and *regularization*, $\Omega(ft)$:

$$obj(\Phi) = \sum_i L(y_i, \hat{y}i) + \sum t\Omega(f_t) \tag{5}$$

The *training loss*, L, estimates the model's goodness of fit based on the training data. A common form of L for classification, which is used in this research, is the logistic loss (i.e., binary logistic) for $y \in 0, 1$ [72]:

$$L_{Logistic} = -\frac{1}{N}\sum_{i=1}^{N}(y_i log(p_i) + (1 - y_i) log(1 - p_i)); , \tag{6}$$

where y_i is the true value, $p_i \in [0, 1]$ denotes the probability prediction, and N is the number of samples. An ideal classifier has a logistic loss close to zero.

In order to prevent the model from becoming too complex, a penalty term, denoted as Ω, is applied to the objective function as follows:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2; , \tag{7}$$

Here, γ controls the penalty for the number of leaves, T, and λ is the parameter for controlling the magnitude of leaf weights, ω, in the decision tree. The purpose of including the regularization term in the objective function is to simplify the model and prevent over-fitting. Learning the tree structure is more challenging than a traditional optimization problem, where you can simply take the gradient. Training all the trees simultaneously is not a straightforward task; therefore, XGBoost uses an additive method that optimizes the learned tree and adds a tree at each step. In the $t - th$ iteration, we need to add the following f_t, which minimizes the objective function:

$$obj^t = \sum^{n} i = 1L(yi, \hat{y}^{t-1}i + ft(x_i)) + \Omega(f_t). \tag{8}$$

This function can be simplified and approximated by the Taylor expansion:

$$obj^t \simeq \sum^{n} i = 1 \left[L(yi, \hat{y}^{t-1} i) + gif_t(x_i) + \frac{1}{2} h_i f^2 t(xi) \right] + \Omega(f_t);, \quad (9)$$

where the functions g_i and h_i, the first and second-order gradient of the loss function, are defined as follows:

$$g_i = \partial_{\hat{y}^{;(t-1)}i} L(yi, \hat{y}_i^{;(t-1)}) \quad (10)$$

$$h_i = \partial^2 \hat{y}^{;(t-1)} iL(y_i, \hat{y}_i^{;(t-1)}). \quad (11)$$

We can rewrite Equation 9 by expanding Ω and find the optimal output value (i.e., weight) $\tilde{\omega}_j$ for leaf j as follows:

$$\tilde{\omega} j = -\frac{\sum i \in I_j g_i}{\sum_{i \in I_j} h_i + \lambda};, \quad (12)$$

where I_j is the instance set of leaf j. Replacing Equation 12 and 7 in 9 gives us the following optimal value of the loss function which is used as a similarity score for measuring the quality of each tree structure:

$$o\tilde{b}j^t = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T. \quad (13)$$

For binary logistic loss function we use for the classification, $g_i = -(y_i - p_i)$ (i.e., the residual), and $h_i = p_i(1 - p_i)$ which can be replaced in Equation 12 and 13.

To simplify evaluating tree structure when adding new branches to the tree (i.e., evaluating the split candidates), a greedy algorithm is used. This algorithm starts from one single leaf and adds new branches to the tree iteratively. Therefore, after the tree splits from a given node, the formula for loss reduction (i.e., gain) is as follows:

$$obj_{split} = \frac{1}{2} \left[\frac{\left(\sum_{i \in I_L} g_i \right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i \right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i \right)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma, \quad (14)$$

where I_L and I_R are subsets of the available observations in the left and right nodes after the split. I is the subset of the available observations in the current node so that $I = I_L \cup I_R$. Moreover, the tree structure will continue to split if obj_{split} is positive or other criteria are met, such as the maximum depth of a tree that users need in XGBoost parameters fine-tuning.

Equation 14 is used for finding the best split at any node and it only depends on g_i, and h_i (i.e., the first and second order gradient) of the training loss and the regularization parameter γ. Therefore, as long as the first and second-order gradient is provided, XGBoost can optimize any custom loss function.

XGBoost performs better than other tree boosting algorithms due to (i) having the regularization term for preventing the over-fitting, (ii) downscaling of each new tree by a constant parameter τ to reduce the impact of a single tree on the final model, i.e., it gives the future trees more space to improve the model while reducing the impact of the current tree. Moreover, (iii) XGBoost supports column sampling, which means each tree is built using a subset of the columns from the training dataset.

In the XGBoost method, the "gain" metric represents the average improvement in the optimization objective (e.g., Gini impurity or entropy) that a feature brings when used in the trees. Features with higher gain values are considered more important for the model, as they contribute more to the model's performance. By examining the gain values for each feature, we can gain insights into the relative importance of different features in our dataset. This can help us better understand our data and potentially identify areas where we might want to focus on collecting more data or refining our features.

REFERENCES

[1] X. Qian, S. V. Ukkusuri, C. Yang, and F. Yan, "Short-term demand forecasting for on-demand mobility service," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 1019–1029, Feb. 2022.

[2] X. Xiong, K. Ozbay, L. Jin, and C. Feng, "Dynamic origin–destination matrix prediction with line graph neural networks and Kalman filter," *Transp. Res. Rec.*, vol. 2674, no. 8, pp. 491–503, Aug. 2020.

[3] P. Krishnakumari, H. V. Lint, T. Djukic, and O. Cats, "A data driven method for OD matrix estimation," *Transp. Res. Proc.*, vol. 38, pp. 139–159, Jan. 2019.

[4] X. Shen, Y. Zhou, S. Jin, and D. Wang, "Spatiotemporal influence of land use and household properties on automobile travel demand," *Transp. Res. D, Transp. Environ.*, vol. 84, Jul. 2020, Art. no. 102359.

[5] A. S. Fotheringham, M. E. Charlton, and C. Brunsdon, "Geographically weighted regression: A natural evolution of the expansion method for spatial data analysis," *Environ. Planning A, Economy Space*, vol. 30, no. 11, pp. 1905–1927, Nov. 1998.

[6] G. Atluri, A. Karpatne, and V. Kumar, "Spatio-temporal data mining: A survey of problems and methods," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–41, Jul. 2019.

[7] J.-J. Lin and T.-Y. Shin, "Does transit-oriented development affect metro ridership? Evidence from Taipei, Taiwan," *Transp. Res. Record*, vol. 2063, no. 1, pp. 149–158, Jan. 2008.

[8] L. Anselin and D. A. Griffith, "Do spatial effecfs really matter in regression analysis?" *Papers Regional Sci.*, vol. 65, no. 1, pp. 11–34, Jan. 1988.

[9] M. Deng, W. Yang, Q. Liu, R. Jin, F. Xu, and Y. Zhang, "Heterogeneous space–time artificial neural networks for space–time series prediction," *Trans. GIS*, vol. 22, no. 1, pp. 183–201, Feb. 2018.

[10] S. Cheng, F. Lu, P. Peng, and S. Wu, "Multi-task and multi-view learning based on particle swarm optimization for short-term traffic forecasting," *Knowl.-Based Syst.*, vol. 180, pp. 116–132, Sep. 2019.

[11] H. Sun, H. X. Liu, H. Xiao, R. R. He, and B. Ran, "Use of local linear regression model for short-term traffic forecasting," *Transp. Res. Rec.*, vol. 1836, no. 1, pp. 143–150, Jan. 2003.

[12] J.-S. Yang, "Travel time prediction using the GPS test vehicle and Kalman filtering techniques," in *Proc. Amer. Control Conf.*, 2005, pp. 2128–2133.

[13] S. Y. Liu, S. Liu, Y. Tian, Q. L. Sun, and Y. Y. Tang, "Research on forecast of rail traffic flow based on ARIMA model," *J. Phys., Conf. Ser.*, vol. 1792, no. 1, Feb. 2021, Art. no. 012065.

[14] D. Xu, Y. Wang, P. Peng, S. Beilun, Z. Deng, and H. Guo, "Real-time road traffic state prediction based on kernel-KNN," *Transportmetrica A, Transp. Sci.*, vol. 16, no. 1, pp. 104–118, Dec. 2020.

[15] Y. Zhang and Y. Liu, "Traffic forecasting using least squares support vector machines," *Transportmetrica*, vol. 5, no. 3, pp. 193–213, Sep. 2009.

[16] R. Wang, D. B. Work, and R. Sowers, "Multiple model particle filter for traffic estimation and incident detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 12, pp. 3461–3470, Dec. 2016.

[17] Y. Qi and S. Ishak, "A hidden Markov model for short term prediction of traffic conditions on freeways," *Transp. Res. C, Emerg. Technol.*, vol. 43, pp. 95–111, Jun. 2014.

[18] Y. Xie, K. Zhao, Y. Sun, and D. Chen, "Gaussian processes for short-term traffic volume forecasting," *Transp. Res. Record*, vol. 2165, no. 1, pp. 69–78, Jan. 2010.

[19] Z. Zhang, "ResNet-based model for autonomous vehicles trajectory prediction," in *Proc. IEEE Int. Conf. Consum. Electron. Comput. Eng. (ICCECE)*, Jan. 2021, pp. 565–568.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.

[22] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, 2017, pp. 1655–1661.

[23] Y. Liang, S. Ke, J. Zhang, X. Yi, and Y. Zheng, "GeoMAN: Multi-level attention networks for geo-sensory time series prediction," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3428–3434.

[24] J. Zhang, Y. Zheng, J. Sun, and D. Qi, "Flow prediction in spatio-temporal networks based on multitask deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 3, pp. 468–478, Mar. 2020.

[25] M. P. H. Raadsen, M. C. J. Bliemer, and M. G. H. Bell, "Aggregation, disaggregation and decomposition methods in traffic assignment: Historical perspectives and new trends," *Transp. Res. B, Methodol.*, vol. 139, pp. 199–223, Sep. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0191261520303489

[26] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph WaveNet for deep spatial–temporal graph modeling," 2019, *arXiv:1906.00121*.

[27] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial–temporal graph convolutional networks for traffic flow forecasting," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 922–929.

[28] A. Li and K. W. Axhausen, "Comparison of short-term traffic demand prediction methods for transport services," *Arbeitsberichte Verkehrs-Und Raumplanung*, vol. 1447, Jul. 2019, doi: 10.3929/ethz-b-000356143.

[29] M. Khalesian, A. Furno, and L. Leclercq, "Improving deep-learning methods for area-based traffic demand prediction via hierarchical reconciliation," *Transp. Res. C, Emerg. Technol.*, vol. 159, Feb. 2024, Art. no. 104410.

[30] X. Fu, G. Yu, and Z. Liu, "Spatial–temporal convolutional model for urban crowd density prediction based on mobile-phone signaling data," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14661–14673, Sep. 2022.

[31] H. Yang, K. Xie, K. Ozbay, Y. Ma, and Z. Wang, "Use of deep learning to predict daily usage of bike sharing systems," *Transp. Res. Rec.*, vol. 2672, no. 36, pp. 92–102, Dec. 2018.

[32] S. Li, X. Liang, M. Zheng, J. Chen, T. Chen, and X. Guo, "How spatial features affect urban rail transit prediction accuracy: A deep learning based passenger flow prediction method," *J. Intell. Transp. Syst.*, vol. 28, no. 6, pp. 1–12, Nov. 2024.

[33] D. Wang, Y. Yang, and S. Ning, "DeepSTCL: A deep spatio-temporal ConvLSTM for travel demand prediction," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[34] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/04df4d434d481c5bb723be1b6df1ee65-Paper.pdf40

[35] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[36] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," 2017, *arXiv:1709.04875*.

[37] Z. Pan, Y. Liang, W. Wang, Y. Yu, Y. Zheng, and J. Zhang, "Urban traffic prediction from spatio-temporal data using deep meta learning," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 1720–1730.

[38] Z. Diao, X. Wang, D. Zhang, Y. Liu, K. Xie, and S. He, "Dynamic spatial–temporal graph convolutional neural networks for traffic forecasting," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 890–897.

[39] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," 2013, *arXiv:1312.6203*.

[40] M. Welling and T. N. Kipf, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016. [Online]. Available: https://bibbase.org/service/mendeley/bfbbf840-4c42-3914-a463-19024f50b30c/file/25dbdd06-4704-a33f-23d9-c626b08adc1e/160902907.pdf.pdf

[41] Y. Rajabzadeh, A. H. Rezaie, and H. Amindavar, "Short-term traffic flow prediction using time-varying vasicek model," *Transp. Res. C, Emerg. Technol.*, vol. 74, pp. 168–181, Jan. 2017.

[42] L. Zhao et al., "T-GCN: A temporal graph convolutional network for traffic prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3848–3858, Sep. 2020.

[43] X. Wang, Y. Shang, and G. Li, "DTM-GCN: A traffic flow prediction model based on dynamic graph convolutional network," *Multimedia Tools Appl.*, vol. 83, no. 41, pp. 89545–89561, Feb. 2024.

[44] H. Yan and X. Ma, "Learning dynamic and hierarchical traffic spatiotemporal features with transformer," 2021, *arXiv:2104.05163*.

[45] S. Zhang, H. Zheng, H. Su, B. Yan, J. Liu, and S. Yang, "GACAN: Graph attention-convolution-attention networks for traffic forecasting based on multi-granularity time series," 2021, *arXiv:2110.14331*.

[46] J. Meppelink, J. Van Langen, A. Siebes, and M. Spruit, "Beware thy bias: Scaling mobile phone data to measure traffic intensities," *Sustainability*, vol. 12, no. 9, p. 3631, May 2020.

[47] Centraal Bureau voor de Statistiek (CBS). (2017). *Gegevens Per Postcode*. Accessed: Feb. 1, 2023. [Online]. Available: https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/gegevens-per-postcode

[48] Y. Seo, M. Defferrard, P. Vandergheynst, and X. Bresson, "Structured sequence modeling with graph convolutional recurrent networks," in *Proc. Int. Conf. Neural Inf. Process.*, 2018, pp. 362–373.

[49] Z. Eftekhar, S. Behrouzi, P. Krishnakumari, A. Pel, and H. van Lint. (2024). *The Codes Associated With the Publication: The Role of Spatial Features and Adjacency in Data-Driven Short-Term Prediction of Trip Production: An Exploratory Study in The Netherlands*. [Online]. Available: https://data.4tu.nl/datasets/51fa919d-bc31-4e55-92ac-6fc67ff50fcc/1

[50] Z. Eftekhar, A. Pel, and H. van Lint. (2024). *The Input Data Associated With the Publication: The Role of Spatial Features and Adjacency in Data-Driven Short-Term Prediction of Trip Production: An Exploratory Study in The Netherlands*. [Online]. Available: https://data.4tu.nl/datasets/835e7093-024c-4bed-b7e5-28f1582f5998/1

[51] C. M. Poteraş, M. C. Mihăescu, and M. Mocanu, "An optimized version of the K-means clustering algorithm," in *Proc. Federated Conf. Comput. Sci. Inf. Syst.*, Sep. 2014, pp. 695–699.

[52] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[53] R. Cohn and E. Holm, "Unsupervised machine learning via transfer learning and K-means clustering to classify materials image data," *Integrating Mater. Manuf. Innov.*, vol. 10, no. 2, pp. 1–14, Jun. 2021.

[54] W. V. Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool, "SCAN: Learning to classify images without labels," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 268–285.

[55] H. Sun, Y. Chen, J. Lai, Y. Wang, and X. Liu, "Identifying tourists and locals by K-means clustering method from mobile phone signaling data," *J. Transp. Eng., A, Syst.*, vol. 147, no. 10, Oct. 2021, Art. no. 04021070.

[56] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-means clustering algorithm," *J. Roy. Stat. Soc. Ser. C Appl. Statist.*, vol. 28, no. 1, pp. 100–108, 1979.

[57] P. J. Lisboa, T. A. Etchells, I. H. Jarman, and S. J. Chambers, "Finding reproducible cluster partitions for the k-means algorithm," *BMC Bioinf.*, vol. 14, no. S1, p. 8, Jan. 2013.

[58] M. I. Shamos and D. Hoey, "Closest-point problems," in *Proc. 16th Annu. Symp. Found. Comput. Sci. (SFCS)*, Oct. 1975, pp. 151–162.

[59] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.

[60] C. Ning and Z. Hongyi, "An optimizing algorithm of non-linear K-means clustering," *Int. J. Database Theory Appl.*, vol. 9, no. 4, pp. 97–106, Apr. 2016.

[61] T. van Elteren, "A comparative study of human engineered features and learned features in deep convolutional neural networks for image classification," Ph.D. dissertation, Fac. Sci. Eng. Comput. Sci., Univ. Groningen, Groningen, The Netherlands, 2018.

[62] T. Kaur and T. K. Gandhi, "Deep convolutional neural networks with transfer learning for automated brain image classification," *Mach. Vis. Appl.*, vol. 31, no. 3, pp. 1–16, Mar. 2020.

[63] C. Wang et al., "Pulmonary image classification based on inception-v3 transfer learning model," *IEEE Access*, vol. 7, pp. 146533–146541, 2019.

[64] V. Iglovikov and A. Shvets, "TernausNet: U-Net with VGG11 encoder pre-trained on ImageNet for image segmentation," 2018, *arXiv:1801.05746*.

[65] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[66] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
[67] J.-H. Hong and S.-B. Cho, "A probabilistic multi-class strategy of one-vs.-rest support vector machines for cancer classification," *Neurocomputing*, vol. 71, nos. 16–18, pp. 3275–3281, Oct. 2008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231208003007
[68] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[69] J. Zhai, J. Qi, and C. Shen, "Binary imbalanced data classification based on diversity oversampling by generative models," *Inf. Sci.*, vol. 585, pp. 313–343, Mar. 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025521011804
[70] A. K. Srivastava et al., "Winter wheat yield prediction using convolutional neural networks from environmental and phenological data," *Sci. Rep.*, vol. 12, no. 1, pp. 1–14, Feb. 2022.
[71] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, Feb. 2002.
[72] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*, vol. 4. Cham, Switzerland: Springer, 2006.

Zahra Eftekhar received the B.Sc. and M.Sc. degrees in civil engineering (specialization in transportation). She is a Ph.D. Researcher with the Department of Transport and Planning, working specifically in the data analytics and traffic simulation lab—DiTTlab. Her research is focused on traffic demand estimation and prediction based on data assimilation and machine learning.

Saman Behrouzi received the M.Sc. degree in information technology–computer networks from Sharif University of Technology, Iran. He was a Research Assistant at the University of Tehran, Iran. He is a Ph.D. Researcher with the Department of Transport and Planning. His main contributions as a research assistant were designing machine learning models and data analysis in different fields of research like bioinformatics, complex network analysis, and bibliometrics. His main focus is on multiscale visualization of different types of transportation networks.

Panchamy Krishnakumari received the double M.Sc. degree in computer science from KTH, Sweden, and TU Delft, The Netherlands, and the Ph.D. degree (cum laude) from TU Delft in February 2020. During her Ph.D., she worked part-time as a Traffic Data Scientist at CGI Netherlands BV for three years. She is an Assistant Professor of data-driven multiscale modeling for traffic and transportation and the Co-Director of the Artificial Intelligence for Mobility Laboratory, Department of Transport and Planning. Her research is on developing interpretable machine learning models for understanding the mobility dynamics of large-scale multimodal networks. Her interest in finding patterns in data led her to the transport domain and to conduct her doctoral research at TU Delft in multiscale pattern recognition of transport network dynamics and its applications.

Adam Pel received the Ph.D. degree in evacuation modelling and management in 2011. He is an Assistant Professor of transport modelling and traffic analyst at Fileradar. He joined Delft University of Technology as an Assistant Professor. Since 2014, he has also been a Research and Development part-time at Fileradar on data-driven traffic analytics and predictions. His main fields of expertise are data analytics, mathematical modelling, and simulation and optimisation techniques for (road) transport systems. His research topics include travel behaviour, the performance of transport networks, and particularly the resilience of transport systems. He has chaired several international conferences and is an Associate Editor for the journals *Transportation Research Part C: Emerging Technologies* and *Transportation Science*.

Hans van Lint received the M.Sc. degree in civil engineering informatics and the Ph.D. degree in transportation from Delft University of Technology (DUT) in 1997 and 2004, respectively. He was an Information Analyst and a Transport Engineer with various organizations. He was appointed as the Anthonie van Leeuwenhoek Full Professor (an honor reserved for only a few young, talented scientists and educators) by the Executive Board, DUT, in 2013. He has co-authored more than 55 peer-reviewed journal articles. His expertise lies at the interface between traffic flow theory and simulation, data analytics, and machine learning techniques. He has worked (with his colleagues and Ph.D. students) on travel time reliability, new traffic flow theories and models, and data assimilation and fusion methods for estimating and predicting the traffic state in networks. Many of these research projects have led to follow-up and valorization projects in which the developed models and tools have led to innovative applications in practice in his laboratory (dittlab.tudelft.nl). He serves as an Associate Editor for IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS and is active in many international projects and collaborations.