Ph.D. DISSERTATION

Predicting the Impact on Speed Reduction in Adjacent Networks of a Link Using the Graph Attention Model

그래프 어텐션 모델을 활용한 링크의 인접 네트워크 통행 속도 감소 영향력 예측

2023년 8월

서울대학교 대학원

공과대학 건설환경공학부

함 승 우

Predicting the Impact on Speed Reduction in Adjacent Networks of a Link Using the Graph Attention Model

그래프 어텐션 모델을 활용한 링크의 인접 네트워크 통행 속도 감소 영향력 예측

지도교수 김 동 규

이 논문을 공학박사 학위논문으로 제출함

2023 년 6 월

서울대학교 대학원

공과대학 건설환경공학부

함 승 우

함승우의 공학박사 학위논문을 인준함

2023 년 6 월

위 원 장	이 청 원
부위원장	김 동 규
위 원	고 승 영
위 원	박 호 철
위 원	김 인 희

Abstract

Predicting the Impact on Speed Reduction in Adjacent Networks of a Link Using the Graph Attention Model

Seung Woo Ham
Department of Civil and Environmental Engineering
College of Engineering
Seoul National University

Traffic congestion has long been recognized as a significant impediment to urban mobility, causing delays, increased travel times, and considerable economic and environmental costs. In light of these challenges, this study aims to identify the influence of links within a road network on adjacent networks to prioritize them for future applications. Focusing on the urban road network of Seoul, South Korea, we developed an impact on adjacent network index and a high-performance prediction model for network-scale speed reduction. The model incorporates the property of traffic flow and heterogeneity of road networks, accounting for interrupted and uninterrupted flows. Furthermore, we introduced a loss function for attention values to enhance their realism and the reliability of prediction results. Consequently, when paired with a graph attention model,

the traffic flow-aware adjacency matrix demonstrated enhanced performance in comparison to the traditional distance-based adjacency matrix. Also, applying the heterogeneity of road networks brought advanced performance in speed reduction prediction tasks. Adding an attention loss weakened the prediction task, which is natural but strengthened the recall of the true data. Our results demonstrate the model's real-time performance and its potential for practical applications in various traffic scenarios. The results of this model are anticipated to be concurrently used in transportation operations such as signal optimization and traffic planning like road expansion.

Keywords: Graph Attention Model, Speed Reduction, Impact on Adjacent Network, Heterogeneous Road Network, Attention Loss

Student Number: 2018-25029

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Roads are vital infrastructure connecting a country's significant bases and play a crucial role in national development. These transportation systems are responsible for transporting resources, such as people and goods, to their required locations, thereby satisfying social needs and enhancing productivity. Consequently, efficient road usage generates economic benefits by reducing driving costs and time. Roads demonstrate efficient transport capacity up to their maximum limit, with traffic volume increasing linearly as new vehicles enter the road. However, traffic congestion occurs if vehicles continue to enter the road beyond its capacity, and transportation efficiency declines sharply. At this point, the upward trend on the density-traffic volume graph ceases, and the traffic volume at the peak begins to fall, allowing for only minimal traffic processing compared to the road's capacity.

Traffic congestion causes significant economic losses due to extended travel

times and energy inefficiency, making it a pressing issue for urban areas. The rapid urbanization and population growth in these areas have increased the number of vehicles on the roads, exacerbating congestion issues. Hence, urban areas became the primary victim of traffic congestion. According to a 2020 study by the Korea Transport Institute on national traffic congestion costs, Seoul's traffic congestion cost amounted to 11.55 trillion won, which increased to 31.05 trillion won when combined with Gyeonggi-do and Incheon urban areas. Although Seoul's population has been declining steadily since 2010, congestion costs continue to rise due to increased commuting distances from suburban commuters (Jun, 2020) and a greater number of vehicles per household.

Countermeasures such as controlling traffic signals are implemented to address traffic congestion to improve transportation efficiency. However, the available resources are insufficient compared to the demand at various points. Additionally, the current allocation of resources is not based on data analysis but on a practical level. When congestion shifts from recurrent to non-recurrent, allocating resources to appropriate locations becomes even more complicated. As the drivers have never encountered the same type of congestion, the response to non-recurrent congestion will likely worsen. This is due to the unique nature of non-recurrent congestion that occurs differently each time. Given the growing frequency and severity of traffic congestion resulting from urbanization (Van Aken *et al.*, 2017) and climate change-induced weather disasters (Dawson *et al.*, 2016), the decision-making process for deploying countermeasures must carefully consider the priority links to improve transportation efficiency.

The priority link decision problem is also related to vehicle route selection. Currently, vehicle route selection is based on individual actions approximating user equilibrium. However, if V2X becomes a reality, driving behavior closer to the social optimum can be achieved by managing the traffic volume of priority

links. Efficient resource allocation and detour strategies can only be implemented when the problem of identifying priority links for both recurrent and non-recurrent congestion is resolved.

Understanding the urban road network can contribute to the increased utilization of road infrastructure by improving road operations and individual route strategies. The ability to identify a priority link in an urban network requires a sophisticated understanding of the network. In this study, we focused on identifying links that significantly impact the speed reduction of adjacent networks, designating these as priority links for further analysis and potential intervention.

Various studies have attempted to predict traffic conditions thus far. Among the numerous traffic prediction studies, traffic speed prediction is the most frequently represented topic (Asif *et al.*, 2013; Min and Wynter, 2011; Wang and Shi, 2013). As the most intuitive aspect affecting road users' experience is travel time determined by traffic speed, prioritizing traffic speed is natural. Moreover, speed as a traffic metric offers versatility and ease of data collection advantages. Consequently, this study has adopted speed reduction as the criterion for identifying priority links within the traffic network. Therefore, this study will employ the latest deep learning-based methodology to select priority links in the urban road network. As a result of this study, the rerouting strategy and countermeasure allocation problem can be addressed. Furthermore, this approach enables better decision-making for infrastructure investment and targeted policy implementation, promoting long-term sustainable urban development.

1.2 Research Purpose and Scope

The primary objective of this study is to identify the link's influence on adjacent networks. By leveraging the magnitude of influence, we can prioritize the links within a network. This approach allows transportation engineers to focus their efforts on the most impactful segments of the network, leading to more effective and efficient traffic management strategies. This study targets complex urban road networks, specifically focusing on the city of Seoul, South Korea, as its data source.

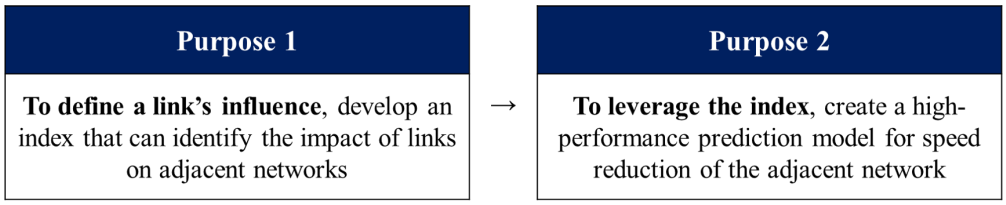


Figure 1.1: The purpose of the research

The First purpose of the research is to develop an index that can identify the impact of links on adjacent networks. The developed index will later be utilized for prediction, ultimately working as a priority index for road networks. The objective of the index is to measure the impact of current road congestion on future networks. When considering the complexity of urban roads, it is a logical fallacy to assume that a specific road's congestion is caused by links that are too far away. Therefore, the number of hops included in the index was appropriately adjusted. (In a graph, a 'hop' refers to the path length between the source and the destination. Two links are considered '1-hop' apart if they are separated by one intersection.)

Next, an engineering technique for an advanced understanding of urban road networks has been proposed. This engineering technique encompasses the

second purpose. The second purpose is to create a high-performance prediction model for speed reduction of the adjacent network.

The adjacency matrix within the graph attention model has been reconstructed as a traffic flow-aware adjacency matrix, which replaces the distance-based adjacency matrix. This traffic flow-aware matrix takes into consideration the direction of traffic flow and congestion propagation. This adjacency matrix overcomes the drawback of the distance-based matrix, which is used widely throughout the literature. By leveraging the power of the adjacency matrix to reflect n-hop connections, the number of layers can be reduced. This enables the identification of relationships between links with fewer layers and allows for faster computation.

Another technique is to reflect the heterogeneity of the road network. Model structure improvement has been made toward modifying the components of the graph attention model to reflect heterogeneity. There are two types of traffic flow: interrupted and uninterrupted flow. Particularly in urban areas, interrupted and uninterrupted flow exist at different levels and form separate road networks. Previous models have lacked consideration for these differences. This study addressed this issue by incorporating these distinctions into the model.

A loss related to the attention value within the model has been added to enhance the realism of road network analysis. The attention value is merely a parameter within the model; as such, the model primarily focuses on predicting speed reduction rather than the interpretability of the attention value itself. However, this approach leads to issues with the attention value's lack of realism, which will be addressed later in this study. After predicting the influence in adjacent networks of each link is completed, operational strategies are recommended at a qualitative level.

1.3 Research Contribution

The contributions of this paper can be summarized into four aspects.

First, a novel adjacent network impact index and its prediction model have been proposed. This index provides a simple description of roads' impact on future networks. Combined usage with a prediction model will enable us to decide the countermeasure locations and strategies against congestion.

Second, a traffic flow awareness structure has implied in the model. The traffic flow goes from upstream to downstream, and congestion propagates from downstream to upstream. Also, a traffic network is a regular network compared with other networks such as social networks. The model has reflected these characteristics of traffic and road networks using an adjacency matrix.

Third, the model's performance is enhanced by considering the characteristics of the heterogeneous road network. Urban roads, consisting of interrupted and uninterrupted flows, possess distinct features that need to be individually addressed. The proposed model improves its understanding of the road network by incorporating road heterogeneity between interrupted and uninterrupted flows.

Fourth, the model induces more realistic attention values, thereby increasing the reliability of prediction results. Attention values exist only as parameters in the model, and the model's learning focuses primarily on prediction accuracy, often overlooking interpretability. This characteristic can severely undermine the realism of the model's internal parameters. We introduced a loss function for the attention values to enhance their realism. The remaining paragraphs in this section describe each contribution in more detail.

First Contribution: The research developed a congestion index on the network side, which has not been implemented before. This index will reveal

the impact of roads on the network that other indices have been unable to show. Moreover, this index is predicted by a deep learning model, so it also has a predictability to unseen conditions. Non-recurrent congestion is often caused by rare events such as natural disasters or accidents. Based on past data, existing statistical models cannot respond effectively to these events. Optimization models have been proposed to address this issue, but they can be computationally complex and lack real-time performance as road network size increases. Furthermore, these models cannot predict future situations in high performance as they predict the future with internal human-made logic.

The results of this study demonstrate the ability to achieve real-time performance that was not possible with the previous priority link identification method. Nvidia's Tesla A100 GPU can process the Seoul road network within 5 seconds. Even if the model is implemented on a device with significantly lower computational power, provided there is sufficient graphic memory to upload it, the model can respond within the 5-minute target time. Additionally, the adjacency matrix, non-linear function, and attention mechanism incorporated in the graph attention model contribute to deeply understanding traffic in urban networks. Consequently, the model can operate robustly across various scenarios. A mere three weeks of data is sufficient for model training, highlighting the model's strengths in practical applications.

Second Contribution: The model incorporates a traffic flow-aware adjacency matrix. Traffic flow progresses from upstream to downstream, with congestion propagating in the opposite direction, from downstream to upstream. This connection-based flow propagation should be considered in a model. Moreover, traffic networks exhibit greater regularity than other networks, such as social networks. These unique traffic characteristics are accounted for within the model by utilizing an adjacency matrix. Through the literature, various mod-

els have been applied to predict traffic speed, and with the emergence of deep learning-based traffic speed prediction models in 2014, prediction accuracy has increased significantly (Zhang *et al.*, 2019b; Wang *et al.*, 2019; Jia *et al.*, 2016). Recently, graph-based deep learning methodologies have emerged, and as road networks themselves are graphs, these methodologies are being actively applied to network traffic speed prediction tasks (Yu *et al.*, 2020; Lu *et al.*, 2020).

Graph-based deep learning methodology has the advantage of reflecting the relationship between multiple data points. Domain knowledge can be involved by feeding the appropriate adjacency matrix, inducing message propagation to follow the purpose. However, a non-parametric method can also be applied using the attention mechanism. The attention mechanism considers the significance of the specific data by comparing the similarity between data sets. The model focuses more on influential data Vaswani *et al.* (2017), which has assigned a high attention value. The attention mechanism allows the model to learn the relationship in a non-parametric way instead of the user-determined relationship between the links. The graph attention model is a model that combines both graph-based methodology and an attention mechanism (Velickovic *et al.*, 2017). By blending the strengths of these two methodologies, they exhibit a synergistic effect in predicting traffic speed (Zhang *et al.*, 2019a; Kong *et al.*, 2020; Zhang *et al.*, 2020; Park *et al.*, 2020; Zheng *et al.*, 2020).

However, existing graph attention model research in traffic speed prediction has lacked consideration of traffic flow. In most papers, the numeric value of the adjacency matrix has been determined by the distance between the data collection points, which are referred to as "links" or "road segments." A short distance between two links does not necessarily mean that this pair highly influences each other. Utilization of the Graph Neural Network (GNN) tailored to the road network structure is also insufficient. When too many layers are

stacked within a GNN, a cycle structure occurs between nodes in a graph. This cycle affects the interconnected nodes and causes the over-smoothing problem in which the predictions of all nodes become similar (Liu *et al.*, 2020; Chen *et al.*, 2020). In addition, as the number of layers increases, the probability of occurrence of diverging or vanishing gradients in the overall structure also increases linearly (Chen *et al.*, 2019; Galimberti *et al.*, 2023). Unfortunately, the road network graph is a form of a grid; thus, it is much more regular than other graphs, such as citation graphs. A graph's higher regularity and low connectivity increase the minimum distance between node points; therefore, road network graphs need more graph neural layers than conventional graphs. Consequently, developing a methodology that can avoid the chronic problems of GNN while reflecting specific traffic characteristics is necessary.

Third Contribution: This study reflects the characteristics of real-world road networks in the model. Road networks, in reality, consist of uninterrupted flow and interrupted flow. Uninterrupted flow corresponds to major arterial roads in urban areas, such as the Gangbyeon Expressway, Olympic Expressway, and Dongbu Expressway. These roads have no traffic signals and higher speed limits. While the number of uninterrupted roads is small, they serve as the central axis in urban areas. Interrupted flow encompasses the remaining roads, which are more numerous but have lower road capacities. These roads typically facilitate movement within sub-areas. Including both types of roads without differentiating between them could potentially degrade the model's performance. Therefore, in this study, we have incorporated modules capable of understanding each road network's characteristics.

Forth Contribution: The fourth contribution aims to address the black-box limitations of deep learning. While deep learning studies typically focus on performance metrics, this study emphasizes internal attention values. Attention

values, which is an internal parameter, often sacrifice their value distribution for the sake of accuracy. However, in this study, we constructed a model that considers the realism of parameter values by building an attention value-based loss function.

As discussed later, the model focuses solely on prediction performance if we do not assign appropriate loss constraints on attention values. In this case, the distribution of attention values becomes unrealistic. For example, it may concentrate all attention values on a single link. Focusing the attention value on a single link might be more beneficial, given the simplicity of the traffic phenomenon. However, the attention value at this point cannot be deemed natural. This phenomenon can be easily observed when training the model without imposing an attention loss.

The remainder of the paper is organized as a literature review, where we focus on conventional priority link detection and graph attention model-based methodology. Then, we describe our model details in the methodology section. The model of each stage and data will be described in detail. Finally, the results and conclusion show the outcome of our analysis and implication directions.

Chapter 2

Literature Review

2.1 Traffic Speed and Congestion Prediction

Traffic speed prediction has evolved as a critical component in the efficient and intelligent management of road networks. It plays a pivotal role in formulating traffic control and routing strategies, minimizing congestion, and improving safety. This paragraph delves into the broad range of methodologies and models utilized in traffic speed prediction research. The discussion traces the trajectory from numerical models to cutting-edge deep learning methodologies, underlining the significance of each approach in progressing the field.

Various models attempt to solve the problem of traffic speed prediction using numerical approaches. One notable study in this regard is Dong et al. (2014) (Dong *et al.*, 2014). Dong et al. (2014) put forward numerical state space models that offer several advantages in the field of traffic prediction. The proposed model incorporates both temporal and spatial data, allowing for the consideration of incoming traffic effects and the propagation of shock waves. Furthermore,

the observation equation in the model utilizes occupancy data to calibrate estimation errors over time. The models developed for predicting network flow rate and time mean speed are based on state space models that account for both congested and non-congested traffic, taking into account spatial-temporal patterns to enhance prediction accuracy and robustness. Unlike Autoregressive Integrated Moving Average (ARIMA) and other time series techniques, these models do not require the variable to be stationary. Moreover, the state space estimation method generates equations with a similar structure for stationary and nonstationary data.

Various machine learning techniques can also be found in a vast amount of literature addressing the problem of traffic speed prediction. Pan et al. (2012) focused on predicting speed in the transportation network of Los Angeles County (Pan *et al.*, 2012). They explored the impact of rush hours and events on speed prediction accuracy, particularly for short-term and long-term averages, even in the presence of infrequent occurrences like accidents. By incorporating historical rush-hour behavior, the researchers significantly improved the accuracy of traditional predictors, achieving a 67% enhancement for short-term predictions and a remarkable 78% improvement for long-term predictions. The study compared the performance of two prediction models, ARIMA and the Historical Average Model (HAM). The analysis of real data revealed that ARIMA outperformed HAM when predicting less than 30 minutes in advance. However, HAM demonstrated superior performance for the longer prediction horizon than ARIMA. This result claims ARIMA is less suitable for long-term predictions exceeding 30 minutes in advance.

Asif et al. (2013) introduced unsupervised learning techniques for analyzing the spatiotemporal performance trends in a large-scale prediction system based on Support Vector Regression (SVR) (Asif *et al.*, 2013). The study revealed the

predictability of traffic speeds differed among roads, and the traditional evaluation indices failed to capture the variations across different time periods. The authors identified that certain roads exhibited consistent performance patterns, while others displayed significant variations in performance over time.

Zou et al. (2015) conducted a comprehensive evaluation of the multi-step prediction performance of three models: the Space-Time (ST) model, Vector Auto Regression (VAR), and ARIMA (Zou *et al.*, 2015). Speed data from five loop detectors in Minnesota is used in the research. To capture the cyclical characteristics of the speed data, hybrid prediction approaches are proposed, which decompose the speed into a periodic trend and a residual part. The periodic component is modeled using a trigonometric regression function, while the residual part is modeled using the ST, VAR, and ARIMA models. The results indicate that the ST model outperforms the VAR and ARIMA models for multi-step freeway speed prediction as the time step increases. It also demonstrates that modeling the periodicity and the residual part separately leads to a better understanding of the underlying structure of the speed data. The proposed hybrid prediction approach effectively accommodates the periodic trends and provides accurate predictions for forecasting horizons exceeding 30 minutes.

Following the advent of deep learning methodology, numerous traffic information prediction studies have adopted deep learning techniques for their tasks. One study by Ma et al. (2015) introduces a novel architecture called Long Short-Term Neural Network (LSTM NN) that effectively captures nonlinear traffic dynamics by addressing the issue of back-propagated error decay (Ma *et al.*, 2015). The LSTM NN demonstrates superior performance in terms of accuracy and stability compared to other dynamic neural networks and parametric/nonparametric algorithms. Another research by Wang et al. (2016) focuses on continuous traffic speed prediction using a deep learning

method called Error-feedback Recurrent Convolutional Neural Network (eRCNN) (Wang *et al.*, 2016). This approach incorporates the spatiotemporal information of contiguous road segments and employs error feedback neurons to address abrupt traffic events. The eRCNN outperforms state-of-the-art competitors in terms of predictive accuracy. Zhang et al. (2019) also propose the Attention Graph Convolutional Sequence-to-Sequence model (AGC-Seq2Seq) for multistep traffic speed prediction (Zhang *et al.*, 2019b). This deep learning framework combines the Sequence-to-Sequence (Seq2Seq) model and graph convolution network to capture the complex temporal dynamics and spatial correlations. The attention mechanism and a newly designed training method are introduced to overcome the challenges of multistep prediction and capture temporal heterogeneity. Numerical experiments demonstrate that AGC-Seq2Seq achieves the best prediction performance compared to benchmark models. Future research directions include integrating traffic flow theories and applying the proposed frameworks to advanced transportation management systems.

Several studies have explored the application of advanced deep learning methodologies in traffic prediction (Polson and Sokolov, 2017; Wu *et al.*, 2018; Ma *et al.*, 2017). Polson and Sokolov (2017) utilized a linear model fitted with L1 regularization and tangent hyperbolic non-linear layers. They confirmed the methodology's effectiveness in anomalous cases, such as during Chicago Bears football games and snowstorm events. Wu et al. (2018) attempted to incorporate the spatiotemporal property using a hybrid model called Deep Nural Network based Traffic Flow (DNN-BTF) prediction model. The periodicity of traffic flow was represented through multiple Convolutional Neural Networks (CNN) based on weekly/daily datasets. By including a CNN in the model structure, it was possible to capture the spatial features of the network. Ma et al. (2017) also employed a CNN to capture spatial features, representing traffic speed as a

single image with road time as the axis and the speed of a specific road and time being expressed as an image pixel value. They confirmed that deep learning methodologies outperformed existing statistical models in each paper.

However, research on speed reduction has been relatively limited, with most studies focusing solely on speed prediction. Moreover, in order to improve travel time prediction accuracy, it is crucial to develop a model that can effectively handle the variability and uncertainty in challenging regions. This targeted approach will enable transportation planners and decision-makers to make more informed decisions.

There also exist studies on congestion itself (Nagy and Simon, 2021; Nguyen *et al.*, 2016; Sun *et al.*, 2021). However, in previous research, they defined congestion as a binary variable and built a model to predict congestion propagation paths. Instead of the severity of congestion on each link, they focused their research on the propagation itself. We can intuitively understand that there are severe levels of congestion. Therefore, there is a need to represent congestion as a continuous value.

Table 2.1: Literatures on Congestion Prediction

Title	Author	Year	Data	Model Type	Congestion Identification	Predictivity
Improving traffic prediction using congestion propagation patterns in smart cities	Nagy and Simon	2021	Sacramento, 614 links	XGBoost	Binary	O
Discovering congestion propagation patterns in spatio-temporal traffic data	Nguyen, Liu, and Chen	2016	Sydney 586 links	Dynamic Bayesian Network	Binary	O
Learning traffic network embeddings for predicting congestion propagation	Sun et al.	2021	Singapore, 1,858 links	GCN-LSTM based model	Binary	O

2.2 Priority Link Identification

The representative approach to Priority Link Identification is the network index-based method. A network can be represented using various types of indices, such as degree centrality, eigenvector centrality, and Katz centrality. These network indices offer diverse characteristics for the nodes within the network. Numerous studies have been conducted based on these indices. Bell et al. (2017) introduced a vulnerability assessment technique based on a capacity-weighted spectral network partitioning strategy. They identified priority network linkages as capacity bottlenecks: network limits with the lowest capacity (Bell *et al.*, 2017). Mattsson and Jenelius (2015) categorized vulnerability assessments into topology-based analyses (e.g., connectivity and capacity vulnerability) and system-based analyses, depending on whether the congestion impact by traffic flows was included (Mattsson and Jenelius, 2015). These attempts were primarily applied when estimating an essential link in a fixed graph. From a fundamental perspective, the simulation-based and optimization-based methods introduced below also can be considered derivatives of network index-based research.

Formulating the priority link selection problem as an optimization is also a common approach. Li et al. (2019) investigated a transportation network recovery strategy for the emergency recovery phase based on an optimization problem. They also proposed two resilience metrics to evaluate recovery rapidity and network performance. The link selection strategy was developed using a genetic algorithm. In this case, the genetic algorithm plays a role in selecting a link with high significance. The evaluation part is based on the optimization model (Li *et al.*, 2019). Yang et al. (2016) established a mathematical model to select a priority link based on travelers' heterogeneous risk-taking behavior. They aggregated the research area to alleviate the computational burden (Yang

et al., 2016). Almost all priority link identification problems work at the bi-level to identify the appropriate priority link and observe the effect in the network accordingly. This was also the case with Yang et al. (2016). A bi-level model was also adopted by Yu, Yang, and Yun (2014). The research was conducted to find a priority link based on the link redundancy index in the first step and the link priority index in the second step (Yu *et al.*, 2014; Gu *et al.*, 2020). As the optimization problem becomes more complex, by adding various variables, the optimization problem eventually reaches a level that cannot be solved in closed form, resulting in a long computational time for a solution.

Simulation-based priority link detection consists of a link selection algorithm and network evaluation. The simulation-based algorithm has predictive power as it searches all possible future scenarios but has a computational cost and time weakness. The other problem is that since the future is predicted according to human-made logic, the model may not function properly in unexpected situations that humans did not anticipate. Gauthier et al. (2018) verified the network's resilience when a disruptive event occurred using resilience stress testing and a dynamic mesoscopic simulator. Furthermore, the most critical link among road networks was selected based on the overall travel cost of the entire network. The time difference for the loss of each link was the criteria. However, this study acknowledged a problem with the ranking. The ranking changes rapidly depending on which metric is used. Additionally, it recognized that unavoidable computational costs occurred during the simulation process. Due to the high demand for computational power, it was explained that it is challenging to utilize in real-time, even in a medium-sized network. In a test using the Paris DIRIF road network, which is a medium-sized network comprising 868 links, the selection of a priority link took over an hour (Gauthier *et al.*, 2018).

Another study used a genetic algorithm with a simulation (Pan *et al.*, 2022). A genetic algorithm selects which link to cut through the generations; the simulation starts without the cut-out link. In this study, it is also essential to determine which metric to evaluate network resilience. The simulation results were evaluated with recovery time and cumulative performance during the recovery. This bi-level model formulation is one of the most common model types for selecting priority links. Recent studies tend to select a priority link based on multiple criteria. Aydin et al. (2018) used multiple criteria to select a strategy, such as centrality in network and road hierarchy as criteria (Aydin *et al.*, 2018). Liu et al. (2019) also brought a similar approach for prioritization problem (Liu *et al.*, 2022)

Table 2.2: Literatures on priority link identification

Title	Author	Year	Data	Model Type	Parametric/ Non-parametric	Predictivity
Resilience-based transportation network recovery strategy during emergency recovery phase under uncertainty	Li et al.	2019	Toy Network	Optimization and Simulation	Parametric	Parameterized
Resilience model and recovery strategy of transportation network based on travel OD-grid analysis	Pan et al.	2022	Chengdu Second Ring	Optimization and Simulation	Parametric	Parameterized
Performance of transportation network under perturbations: Reliability, vulnerability, and resilience	Gu et al.	2020	Sioux-Falls Network (Toy)	Optimization and Simulation	Parametric	Parameterized
Prioritizing transportation network recovery using a resilience measure	Liu et al.	2019	Switzerland	Index-based Regression	Parametric	Parameterized
Framework for improving the resilience and recovery of transportation networks under geohazard risks	Aydin et al.	2018	Sindhupalchok road network	Index-based Regression	Parametric	Parameterized

In order to more accurately assess and manage transportation networks, it is crucial to develop new indices or methodologies that can capture the dynamic nature of traffic flow and congestion. By incorporating real-time traffic data and considering the relationships between connected links, these new approaches can provide a more comprehensive understanding of network performance, allowing for more informed decision-making in traffic management and infrastructure planning.

In optimization and simulation modeling, the inherent iterative processes can aptly depict the nuances of road networks, with correlations between roads being illuminated through appropriate model design. While optimization allows for the identification of priority links exclusively in the present context, its confluence with simulation models facilitates prospective predictions. A priority link identification model that can perform prediction can be built with a scheme that renders the future situation through the simulation model and solves the optimization problem for that particular situation. However, optimization and simulation models also have their drawbacks.

Primarily, the computational demands of both optimization and simulation are extensive, precluding real-time determination of priority links. To accurately represent impending traffic conditions within a desired timeframe, simulations must encompass the entire duration. If the interlude between simulation steps is overly extended, it compromises both the utility and precision of the application. Conversely, a shortened time interval necessitates an extensive iterative process, prolonging the attainment of anticipated traffic scenarios. The literature confirmed that the computation time exceeded 1 hour, even in the size of a small village. In the case of deep learning, parallelization using GPUs is well constructed so that many matrix operations can be processed almost instantaneously. However, simulations based on mathematical models are not yet

parallelized through GPU.

Secondly, the simulation model does not accurately predict future scenarios. This issue is a common problem inherent to all simulation models. By approximating real-world situations through finite units, errors are inevitably generated. These errors accumulate over time, resulting in the simulation model's low performance in predicting distant future scenarios. Furthermore, the simulation designer's bias may be reflected in the simulation, causing the overall results to be skewed. Recently, deep learning has been introduced to address this issue, and if the graph attention mechanism can accurately reproduce all traffic situations, precise results can be obtained. However, the model will vary depending on the traffic conditions that the researcher deems most important. The prescribed methodology for addressing this issue employs deep learning techniques with the graph attention mechanism.

Third, different outcomes are obtained depending on the index to be optimized. Many of the performance metrics identified in the literature are determined at the researcher's discretion. These metrics need to be strictly defined since they influence the conclusion of the optimization problem. Nevertheless, akin to the challenges encountered in simulation, a certain degree of human-induced bias remains inevitable.

Fourth, the accumulated data cannot be utilized effectively. Since the future situation is implemented using a predetermined model, newly collected data cannot be incorporated into the model. The model developer should directly modify the internal structure of the simulation to incorporate new data. Until such modifications are executed, the model remains incapable of reflecting the continuously accrued data.

Table 2.3: Existing models and their evaluation with various criteria

Criteria	Index-based	Optimization and Simulation
Real-time	△, May not be possible to increase the network size	X, Real-time unavailable from medium-sized networks
Network Topology	△, Reflected by index, but depends on the intention of the model developer	O, Reflect on the simulation
Link-Link Attraction	△, Reflected by index, but depends on the intention of the model developer	△, Reflect in simulation, but depends on the intention of the model developer
Dynamic property of Traffic	△, Varies by the formulation of the index	O, Reflected as an internal mechanism
Future Prediction	X, Index calculation based on historical data	△, Reflect in simulation, but Depends on the intention of the model developer

2.3 Link and Network Indices

Various congestion metrics are used to evaluate traffic conditions, including speed, travel time, and delay. These metrics provide valuable insights into the performance of individual road links and can help transportation planners and engineers identify problem areas and prioritize improvement projects. However, while these metrics offer a detailed understanding of congestion levels on individual links, they may not fully capture the broader network dynamics and the relationships between connected links (Afrin and Yodo, 2020).

In order to better understand the overall traffic flow within a network and its impact on congestion, it is essential to consider the interdependencies between connected road links. Traditional congestion metrics, mostly a metric for individual links, can overlook the cascading effects of congestion on adjacent links and the more extensive transportation network (Li *et al.*, 2019). The specific limitation of indices can be summarized below

Firstly, certain indices are grounded in linear models, which fail to encapsulate the dynamic nature of traffic. Within traffic flow, vehicles interact with those both in front and behind them. Similarly, vehicular platoons respond to preceding and succeeding platoons. Additionally, when a shock wave traverses the roadway, it too induces interactions. Compared to the traffic response with such a dynamic and non-linear relationship with various elements, the linear model has an inherent problem: it cannot reflect such non-linearity.

Second, some indices do not reflect the topology of the road network. While it is possible to consider the traffic conditions of adjacent roads, quantifying the degree of their interdependence remains a challenge. Therefore, the more complex the road network, the lower the indices' performance.

Third, the indices do not reflect the hierarchy between the two roads. Not all roads are the same; various hierarchies exist among them. Occasionally, some studies take road hierarchy into account. However, even those studies fail to define the relationship between two distinct hierarchies of roads. There may be more important roads among the connected roads, but such characteristics are not taken into account in the existing indices.

Fourth, when the size of the network increases, real-time priority link identification is impossible. If the network consists of 100 links, only 10,000 comparisons corresponding to a 100 by 100-matrix are needed to determine important links. However, if there are more than 5,000 links, like the Seoul road network

conducted in this study, the number of comparisons increases to 2,500,000. In this case, if the comparison for one pair takes longer than 0.1 ms, the prediction for a single step takes more than 5 minutes, making it difficult to utilize at the traffic operation level.

Fifth, the indices cannot have predictive power. As the indices are based on past data. Therefore, it is possible to interpret only past relationships that have already passed. Consideration of the future circumstances to come has not been carried out. This can be a serious problem in real applications. In the case of rare events, such as severe accidents or unprecedented disasters like the heavy rain that occurred in August 2022, there is no comparable historical data available. Even if it is not a rare case, some existing indices cannot interpret the phenomenon unless the exactly same event has occurred in the past, even if it is a frequent event.

Hence, as we can check through the literature, network indices, such as betweenness centrality, degree centrality, Katz centrality, closeness centrality, and eigenvector centrality, have been widely used to analyze and understand the structural properties of transportation networks. These indices help identify important nodes or links within the network and can provide valuable insights for transportation planning and management. Degree centrality represents how directly connected a link is to other links. Katz centrality indicates how many different paths can reach other links. Closeness centrality measures how close the distance is from other links to the target link. Betweenness centrality determines whether a specific link is part of the shortest path between two other links.

However, these indices also have limitations, primarily related to their inability to capture the dynamic nature of traffic within the network. One major limitation of these centrality measures is that they are based on the static structure of the network. When the structure of the network is fixed, the val-

Table 2.4: Definitions of each centrality

Centrality	Definition
Degree Centrality	$C_D(i) = \deg(i)$
Katz Centrality	$C_K(i) = \sum_{n=1}^{\infty} \sum_{j=1}^{N} \alpha^k \left(A^k\right)_{ji}$
Closeness Centrality	$C_C(i) = (N - I) / \sum_j d_{ij}$
Betweenness Centrality	$C_B(i) = \sum_{j \neq i \neq k} \sigma_{jk}(i) / \sigma_{jk}$

ues of these indices are fixed as well, regardless of the actual traffic conditions. Consequently, these static characteristics do not account for the dynamic fluctuations in traffic flow and congestion that are commonly observed in real-world transportation networks. As a result, relying solely on these traditional network indices may lead to an incomplete understanding of traffic patterns and their impact on network performance.

2.4 Graph Attention Model

As deep learning methods demonstrated superior achievements compared to conventional statistical models, graph-based deep learning methods, such as Graph Neural Networks (GNN)s, also showcased their technical value. GNNs are a distinct variant of neural networks specifically designed to handle data structured as graphs. They possess the ability to discern the entire topological configuration of a graph, concurrently updating the properties of both its nodes and edges based on the characteristics of their adjacent elements. This procedure includes the enhancement of node embeddings via a process of aggregation and combining at each respective layer.

$$f_{\mathcal{N}(v)}^{(l)} = \text{AGGREGATE}^{(l)}(\{h_u^{(l-1)}, \forall u \in \mathcal{N}(v)\}) \tag{2.1}$$

$$h_v^{(l)} = \sigma(W^{(l)} \cdot \text{COMBINE}(h_v^{(l-1)}, f_{\mathcal{N}(v)}^{(l)})) \tag{2.2}$$

In Equation 2.1 $h_v^{(l)}$ denotes the feature vector of the node v at the l-th layer, and $\mathcal{N}(v)$ stands for the group of neighboring nodes to a specific node v. At every layer l, a differentiable function part, AGGREGATE, amasses the representation vectors of neighbors, which are then assimilated via the COMBINE function. Furthermore, a weight matrix W is applied, and a nonlinear activation function σ is utilized to refresh the hidden depiction of node v. This model is typically referred to as the message-passing scheme.

The Graph Convolutional Network (GCN) (Kipf and Welling, 2016) is an effective variant of CNN adapted for graph structures. It's a fundamental type of message-passing neural network, using a local neighborhood assembly with first-order spectral filters that are learned, followed by a nonlinear activation function to construct node representations.

$$h_v^{(l)} = \sigma(W^{(l)} \cdot \text{MEAN}\{h_u^{(l-1)}, \forall u \in \mathcal{N}(v) \cup \{v\}\}) \tag{2.3}$$

Based on this GNN, and mostly GCN, numerous research has emerged. Yu, Lee, and Sohn (2020) established an adjacency matrix that considered road length and lane number and inserted a learnable parameter inside the adjacency matrix value, generating a similar effect to the attention mechanism (Yu *et al.*, 2020). Lu et al. (2020) concurrently applied Long Short-Term Memory (LSTM) and GNN (Lu *et al.*, 2020). They used the Xi'an and Beijing feature graphs from road traffic networks and applied the obtained features to LSTM.

Several studies presented attempts to extract spatial and temporal features of the road traffic network using graphs. Ge et al. (2019) utilized k-order spec-

tral graph convolution to approximate the message-passing scheme of a graph (Ge *et al.*, 2019). Using a dilated causal convolution, they constructed a spatiotemporal dependency of road traffic data. Furthermore, the day of the week, road structure, and points of interest were incorporated to advance the model. Li et al. (2021) created a graph where two model substructures were fused simultaneously to create spatial and temporal dependencies. Local and global dependencies were obtained from gated dilated networks (Li and Zhu, 2021). However, these approaches have limitations, as excessive intervention from researchers is needed. Human-made adjacency weights induce the model to be human-dependent.

GNN models had exceptional forecast accuracy; however, they statistically estimated traffic's spatial dependencies, overlooking the possibility of dependencies changing over time. Moreover, the interpretability of deep learning models is insufficient due to their black-box nature. Therefore, a deeper understanding of the road traffic network interdependence derived from the deep learning model is essential.

The attention value captured by the Graph Attention Network (GAT) can represent structural dependencies, providing a higher understanding of the road traffic network and the model. (Researchers have the discretion to adopt the GAT framework when integrating attention mechanisms into graphs. Numerous variations exist.) Several attempts have been made to apply the attention mechanism to graph neural networks without following the GAT framework—these attempts aimed at capturing spatial and temporal attention, respectively. Wang et al. (2020) combined traffic information on adjacent roads with a positional attention mechanism, and a similar approach was taken by Zhou et al. (2021) by reflecting temporary attraction using temporal attention Wang *et al.* (2020); Zhou *et al.* (2020).

The implementation of the GAT framework in traffic information prediction commenced with the work of Zhang et al. (2019). Zhang et al. (2019) integrated the LSTM layer into the GAT framework and utilized the distance between links as the basis for the adjacency matrix (Zhang *et al.*, 2019a). Nevertheless, given that the inception of the attention mechanism aimed to deviate from the existing recurrent neural network (RNN), it is challenging to argue that the combination of LSTM with GAT is a suitable approach. Kong et al. (2020) employed both a self-adaptive adjacency matrix and a distance-based adjacency matrix to augment the non-parametric nature of their model Kong *et al.* (2020). Additionally, they utilized a residual architecture to facilitate information flow across layers. Capturing spatiotemporal features has been a crucial aspect of the GAT framework. Zhang et al. (2020), Zheng et al. (2020), and Park et al. (2020) introduced various layers to capture spatiotemporal features (Zhang *et al.*, 2020; Park *et al.*, 2020; Zheng *et al.*, 2020). Park et al. (2020) enhanced the adjacency matrix by considering connectivity and edge weight, such that two directly connected links are deemed to have connectivity.

Despite the extensive literature, three primary gaps remain. Firstly, prior studies did not construct the adjacency matrix with traffic flow as the focal point. Park et al. (2020) and Yu, Yin, and Zhu (2017) attempted a connectivity-based matrix; however, it concentrated on the link's physical connection rather than the connection established via traffic flow (Park *et al.*, 2020; Yu *et al.*, 2017). Other studies determined the adjacency matrix based solely on distance. This means the consideration of road network topology was insufficient. Insufficient consideration can also be found in other aspects. Numerous papers have employed residual connections; however, the introduction of residual connections was due to the performance demonstrated in other studies, not because of the consideration of traffic-related characteristics (He *et al.*, 2016).

Table 2.5: Literature on traffic state prediction using graph attention model

Title	Author	Year	Data	Adjacency matrix	Purpose
Spatial-Temporal Graph Attention Networks: A Deep Learning Approach for Traffic Forecasting	Zhang, Yu and Liu	2019	PeMSD7	Distance based	Speed Prediction
STGAT: Spatial-Temporal Graph Attention Networks for Traffic Flow Forecasting	Kong et al.	2020	METR-LA PEMS-BAY	Distance based	Speed Prediction
Spatial-Temporal Convolutional Graph Attention Networks for Citywide Traffic Flow Forecasting	Zhang et al.	2020	BJ-Taxi NYC-Taxi NYC-Bike-02	Distance based	Demand Prediction
GMAN: A Graph Multi-Attention Network for Traffic Prediction	Zheng et al.	2020	Xiamen PeMS	Distance ased	Speed Prediction
Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting	Guo et al.	2019	PeMSD4 PeMSD8	Distance based	Speed Prediction
Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting	Yu, Yin and Zhu	2017	BJER4 PeMSD7	Connection and Distance based	Speed Prediction

Secondly, There was no consideration for the various hierarchies of roads present in traffic. All roads were treated as having the same hierarchy, and as a result, no distinction was made between interrupted and uninterrupted flows. The type of road, determined by the presence or absence of signals, significantly influences driving behavior and, therefore, must be taken into account.

Lastly, the examination of attention values was insufficient. There has been no investigation into which roads are deemed essential by each attention value and the effects that emerge accordingly. Although there are examples of case studies conducted at the level of several dozen links, it is challenging to find such literature at a more extensive network scale. In this paper, we will establish a method for obtaining realistic attention values and verify the effects of these attention values on a city scale. This study aims to propose solutions that address these limitations.

The remainder of this section aims to validate the aforementioned limitations via case studies. Numerous research endeavors depict the adjacency matrix of road networks based on proximity. Nonetheless, this methodology presents several logical inconsistencies. It is not accurate to unconditionally connect two links simply because they are nearby, nor is it accurate to refrain from connecting them solely due to a significant distance between them. Figure 2.1 illustrates such a situation. This figure provides a detailed view of Mangwon Hangang Park, revealing that Gangbyeonbuk-ro and Mangwon-ro are situated close to each other. With a straight-line distance of approximately 200 meters, most studies consider these roads interconnected.

However, in reality, accessing Mangwon-ro from Gangbyeonbuk-ro requires a driving distance of more than 2 kilometers, traversing Tojeong-ro and Wausan-ro. This phenomenon is particularly prevalent in the vicinity of urban highways. Olympic Expressway, Gangbyeon Expressway, and Naebu Expressway all pass

by branch roads but are not directly connected to them. As a result, the urban expressway, which plays a crucial role in the road network, generates biased predicted values due to numerous branch roads. Given the importance of urban highways, this issue presents a significant challenge.

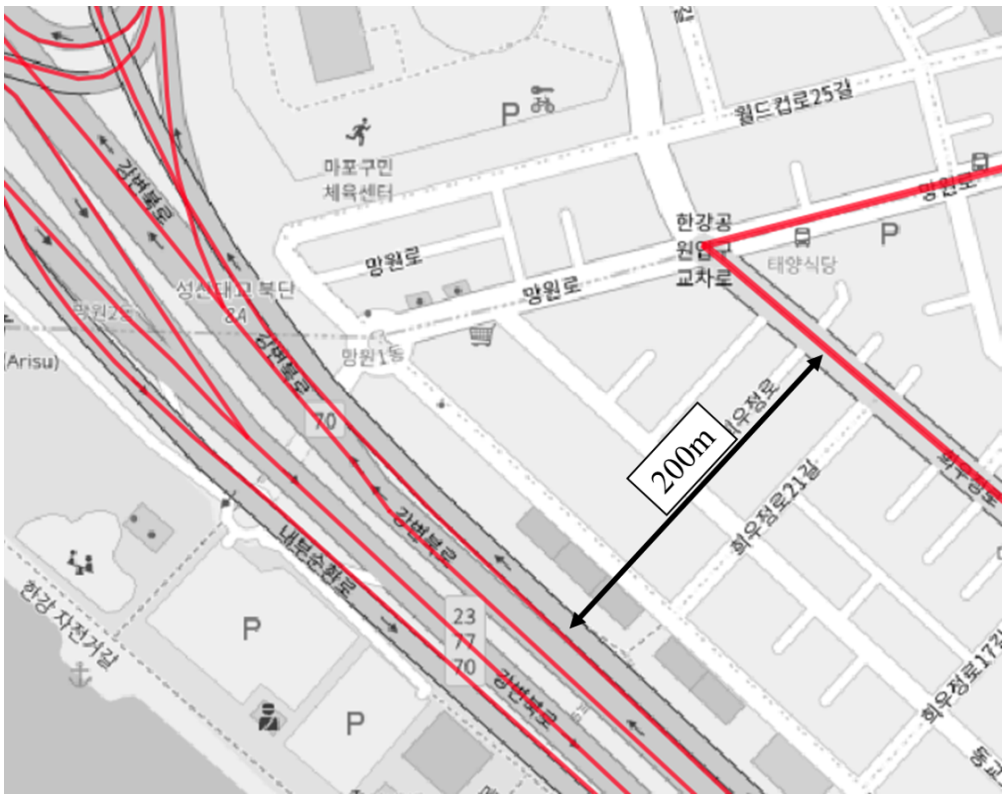


Figure 2.1: Failure case of a distance-based adjacency matrix

An adjacency matrix can be constructed based on connectivity to address this issue. However, this approach introduces another problem stemming from the unique characteristics of road networks, such as that of Seoul, when compared to other types of networks. These unique characteristics can be identified through various graph indices. Table 2.6 presents a comparison of network indices for the Seoul road network, airline data, and authorship data. As the most

representative network type, airline data exhibit a high average degree and low average path length. Its modularity is 0.245, which is lower than that of other graphs. The authorship network has the lowest average degree, at 3.451, and likely due to the nature of thesis writing activities, which are not expected to yield as high a degree as other networks. Consequently, it demonstrates a low density of 0.002 and a high modularity of 0.955.

However, the average path length is the most significant difference between the Seoul road network and the airline and authorship networks. The airline network has an average path length of 2.318, meaning that other airports can be reached with fewer than three stops on average. Despite its low average degree, the authorship network also has an average path length of 5.823, indicating that all individuals can be connected within six hops. In contrast, the average path length of the Seoul road network is 16.73, signifying that, on average, 16.73 movements are required to reach another link. This discrepancy arises from the nature of road networks, which lack a central hub. The absence of a hub further exacerbates the difference between the minimum and maximum path lengths.

The degree distribution depicted in Figure 2.2 also illustrates how the Seoul road network differs from other networks. Since the airline network consists of several hubs and mostly spokes, the degree values are generally low. As the degree value increases, the count consistently decreases. In contrast, the degree distribution of the Seoul road network peaks at the average degree value. Very few links serve as hubs, and even those have degree values that do not differ significantly from other links. As demonstrated in Table 2.6 and Figure 2.2, the Seoul road network represents a highly unique type of graph. Consequently, the performance of the model is limited when approached using the same methods as existing graph attention models.

Finally, directly applying the graph attention mechanism to traffic predic-

Table 2.6: Indices of various networks

Type of Network	Seoul Road Network	Airline Network	Authorship Network
Nodes	5,068	235	1,589
Edges	27,957	1,297	2,742
Average Degree	5.582	11.038	3.451
Network Diameter	46	4	17
Average Path length	16.727	2.318	5.823
Density	0.001	0.047	0.002
Modularity	0.899	0.245	0.955
Average Clustering Coefficient	0.122	0.652	0.878

tion risks overfitting. The attention mechanism was originally introduced to handle challenges in domains such as natural language processing. In natural languages, the vast lexicon can be arranged in myriad combinations to construct sentences. For instance, the term "hard" can denote "difficult" or "exhausting," and simultaneously convey "absolute" or "undeniable." To address such complexities, GPT-4 employs a staggering one trillion parameters.

In contrast, road network data is relatively simple. The speed range is fixed, with no significant deviations from that range. Although there are instances where the speed drops suddenly, it always remains a positive value. As shown in Figure 2.2, the speeds of roads tend to move in tandem. The speeds of links in the road network rise and fall together. They decrease during morning and afternoon rush hours and increase throughout the day, with high speeds guaranteed in the early morning and late at night. Even with a straightforward

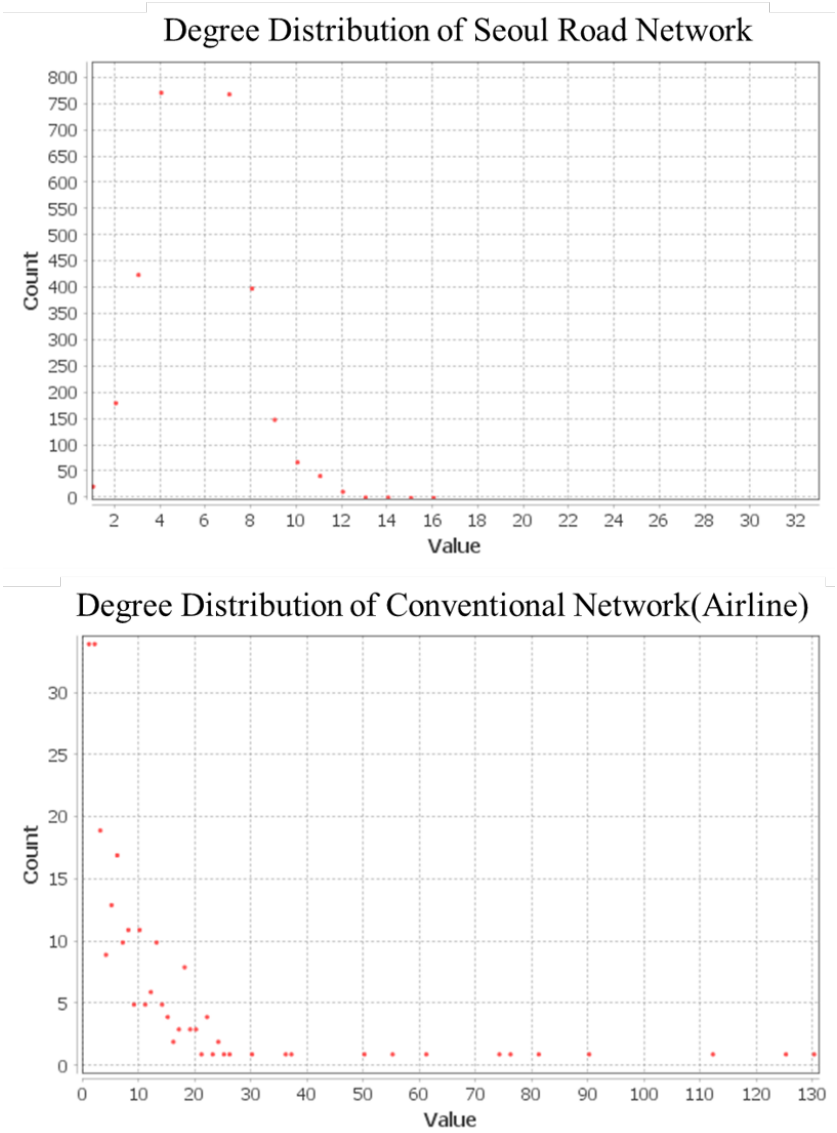


Figure 2.2: Degree distribution of Seoul road network(top) and conventional network (Airline) (bottom)

rule, a range of speeds can be easily predicted. If a model equipped with a large number of parameters used in natural language processing systems is employed

for this simple behavior, overfitting naturally occurs.

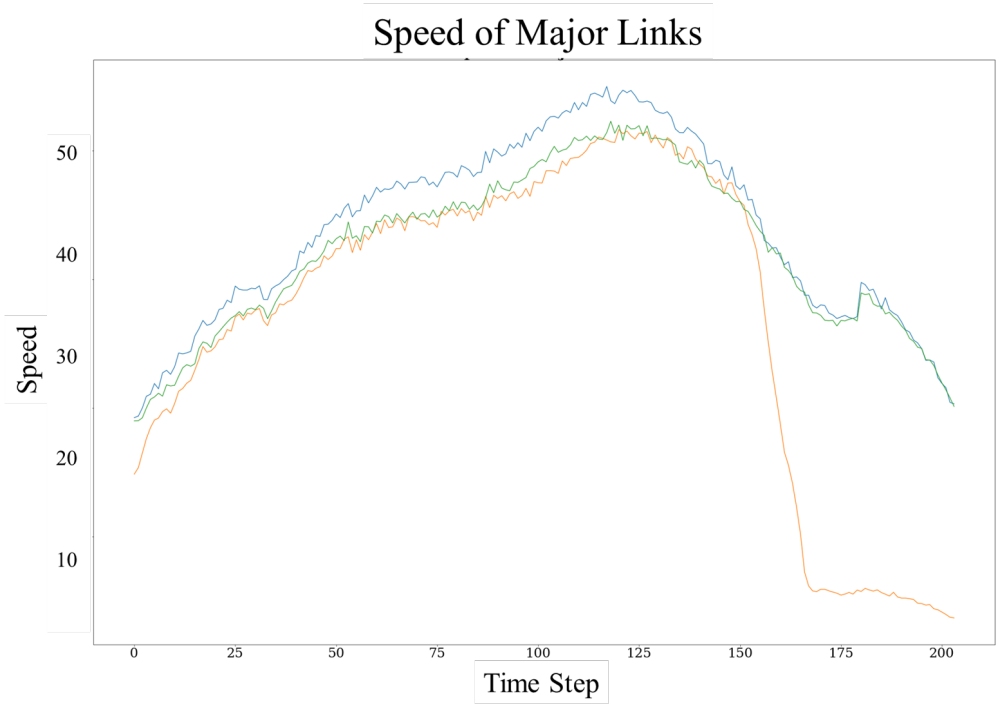


Figure 2.3: Speed of major links showing similar behavior

In particular, the issue becomes more significant since we are also interested in the attention value. While indicators related to speed reduction are primary, this study will also examine whether outliers in attention values occur. Accurate speed predictions and the emergence of biased attention can coexist. In fact, this occurred in numerous cases during the model training. All attention values tend to have the same value or be randomly concentrated on two or three links. All of these problems originate from overfitting. Precisely, as the number of links increases, so does the number of parameters, leading to more severe overfitting. Although the speed reduction prediction performance improves, the attention value already produces results beyond common sense. In the case

of the graph attention model, it was observed that overfitting occurs rapidly as the graph layer - attention - graph layer - attention pattern is repeated. Since we are tackling a problem that has bounded values with a synchronized trend, the attention value is prone to deviate from common sense. It is because the problem could be solved with just the graph layer without an attention mechanism. In the next chapter, which introduces Methods, we will discuss how current research addressed the aforementioned problem.

Chapter 3

Establishment of Impact on Adjacent Network Index

3.1 Index Setup

3.1.1 Research Flow and Data Description

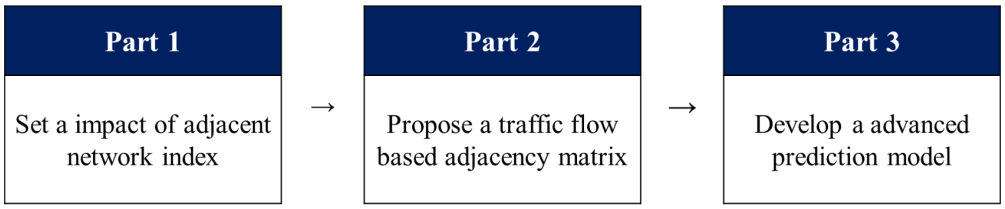


Figure 3.1: The framework and flow of the research

This study mainly consists of three steps. Firstly, we develop an index that can determine the impact of the adjacent network. Secondly, we improve the adjacency matrix of the model to be utilized for predicting the index. Lastly, we develop a prediction model for estimating the index.

In part 1, we define the impact of the adjacent network index. So far, most

deep learning applications in traffic have primarily focused on predicting travel speeds. While travel speed is a crucial factor, this study aims to measure a particular road's influence on the speed reduction of the network.

The index representing the degree of impact on the network can be defined in various ways depending on the researcher's purpose and the given data. The objective of this study is to understand the impact a specific road has on the degradation of travel speeds within the network. Accordingly, the research was conducted using base indicators related to speed. The new index is the time-space combination of existing indicators.

In some studies, attention values are directly used to assess the importance of links. While it is possible to consider attention values as direct indicators of importance, current research chooses not to do so. This decision was made because the validation of attention values has not yet been sufficiently conducted, and developing a direct indicator is more straightforward.

Instead of measuring the speed itself, we measured the degree of speed reduction. The appropriate speed can vary for each road. Even if vehicles pass at the same speed of 15 km/h, the congestion level will have different values depending on the road's appropriate speed. For instance, if the link speed is measured at 15 km/h, the speed reduction ratio for a road with an appropriate speed of 30 km/h would be 50%, while for a road with an appropriate speed of 60 km/h, the ratio would be 75%. Based on such speed reduction indicators, we aimed to investigate the impact on the degree of speed reduction in the network.

In Part 2, we improved the shortcomings of the existing graph attention model by using data from Gangnam-gu. The fields in which the graph attention model has been actively developed are social network services, recommendation systems, and pharmacies. These fields have few things in common with trans-

portation. For this reason, there are model characteristics that are not suitable for the characteristics of traffic. The most significant difference when comparing the example with the social network service is the absence of a hub in the social network. There may be routes that act as significant axes due to high traffic volume, but according to the road network data, there is almost no difference in the node degree of the main routes and the branch lines. One problem that occurs because of this is that the maximum distances between nodes are lengthy. In Part 2, this problem was solved by improving the adjacency matrix. Details on this are described in the problem definition, which will be described later.

This research improved the model based on data from Gangnam-gu data because of the practical computational cost problem. The time difference between training and testing is significant in deep learning models. In the case of training, the gradient of each parameter identified in the loss should be updated every batch. Since deep learning is a model with enhanced explanatory power by using a large number of parameters, the update process takes a considerable amount of time. On the other hand, in the case of testing, there is no need to store the gradient; only checking the output is necessary, thus taking much less time. In this study, when training was conducted throughout Seoul, it generally took about 10 hours for the train and validation errors to converge. Therefore, improving the model structure using this data may take too long.

The structure of urban roads in cities is grid structures, which are generally similar. Therefore, it is expected that there will be no significant problems even if the adjacency matrix is developed based in Gangnam-gu, one of the most urbanized areas in Seoul.

Lastly, in Part 3, the index developed in Part 1 and the adjacency matrix made in Part 2 were used for the entire city of Seoul. In this part, we sought to incorporate traffic characteristics into the model. There are two main types of

traffic flow: interrupted flow, which has traffic signals, and uninterrupted flow, which has no traffic signals. These two flows possess different hierarchies on the road and exhibit distinct characteristics. Uninterrupted flow typically has a higher speed limit and broader road width. In contrast, interrupted flow exhibits opposite properties: lower speed limit and narrower road width. Although it is clear that the hierarchy of different roads should be treated separately, there have been no attempts to the best of our knowledge. In this study, we aim to develop a model that reflects the hierarchy of roads with varying characteristics.

In addition, this study introduced a loss to limit unrealistic attention values. Previous analyses of attention values have not been adequately conducted, particularly in network-scale studies like this one. Ideally, attention values should be well-distributed, referencing links close to a specific link. However, as the number of referenced links increases, attention values begin to deviate in unexpected directions. This is expected to have a relationship with the synchronized behavior of the traffic state.

Most of the speed of a city shows a similar pattern. During the morning and afternoon peak hours, the overall travel speed decreases; at other times, the travel speed increases. In other words, from the point of view of the graph attention model, even if the weight of the attention score is adjusted less sensitively, the model can respond appropriately. In reality, the attention value is focused on one single link.

The sum of attention each link can refer to equals one. Therefore, if you check the maximum attention of each link, you can find out which road is considered the most important. The problem is that most roads give maximum attention to similar or nearly identical roads. There may be various reasons for this, but the simplicity of the traffic speed data described above is thought to play a major role. For this reason, human intention was included in the

model development stage. The sum of each road's attention scores and distance penalty was checked for both uninterrupted and interrupted flow. Here the distance penalty was given if the referred link was located too far.

The rest of this section describes the data. In Seoul, the traffic information collection agency provides traffic data through Transport Operation & Information Service (TOPIS). Information such as link speed, traffic volume, and public transportation usage is provided. Of these, link speed data was used in this study. Traffic volume data were excluded because the collection location was less than 3% of the link speed, and there were too many missings.

The link speed provided by TOPIS is recorded based on the service link. However, there is a difference between the standard node-link system and the service link system provided by the National Traffic Information Center. TOPIS provides mapping data between the service and standard links to solve this problem. In general, one service link consists of 1 to 4 standard links. However, given the sheer volume of links, missing elements in the mapping data are inevitable. The number of standard links obtained based on the Service link-Standard link mapping data was counted as 11,398. Considering that there were 24,720 existing mapping data, more than half were lost due to missing. Since one of the essential ideas of this study is the improvement of the adjacency matrix, this loss of connectivity adversely affects the study results. Therefore, if there is a service link with speed data within 3-hop, the link has been restored. In this way, the number of repairing a broken standard link corresponds to about 2,500. Therefore, the total number of service links used in the Seoul study was counted as 5,068. Figure 3.2 shows the illustration of the service links in Seoul.

The acquisition period was from November 1 to November 28, 2022, and data were acquired for 28 days. The unit of aggregation of the data is 5 minutes. It is confirmed that as social distancing restrictions in Seoul were gradually

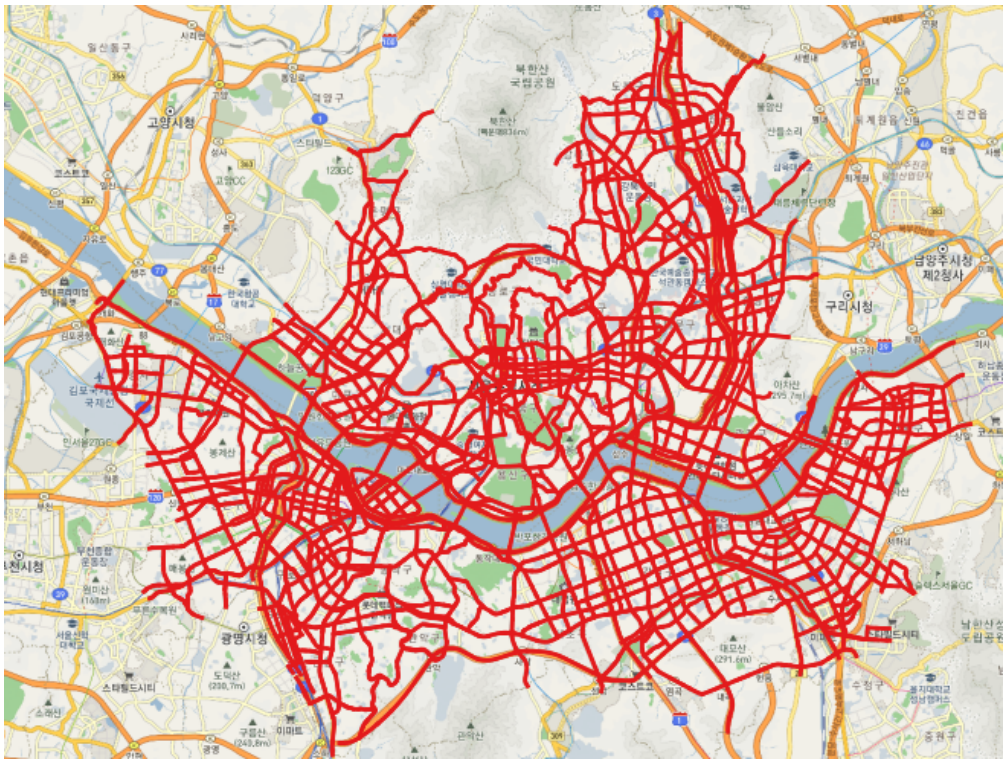


Figure 3.2: The service link of Seoul used in the research

lifted in September 2021, the transportation demand in November 2022 would have been equal to that of a typical year. Since some companies still recommend working from home, slight differences may exist, but it is expected to be minimal.

Gangnam-gu data used in Part 1 consists of 228 links selected from Seoul data. Using these 228 links, we proceeded with model improvement. All major arterial roads in the vicinity of Gangnam-gu are included in this data.

3.1.2 Utilizing Speed Reduction Index

There are various indicators for measuring the state of a network. We can obtain various indicators by combining primary traffic data such as speed, density,

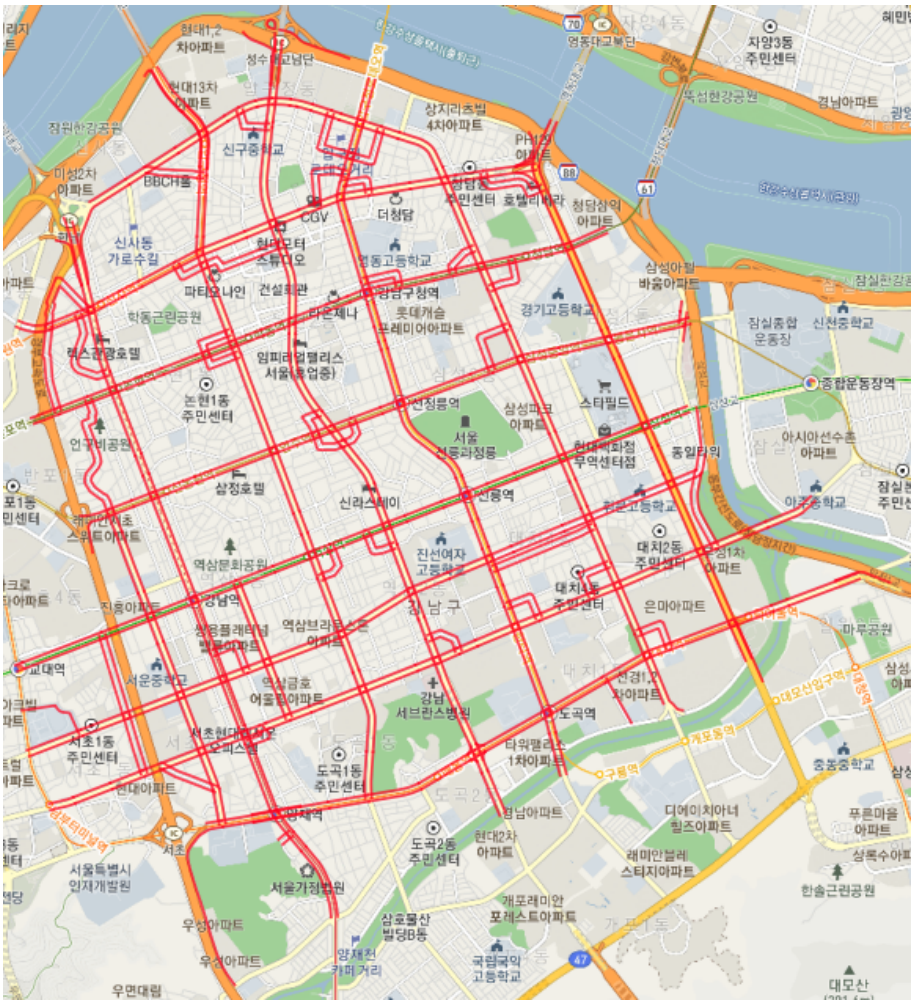


Figure 3.3: Target service link of Gangnam-gu used in the research

and volume. Speed is the simplest, most stable, and easiest to acquire among the various fundamental indicators. Although data acquisition points for traffic volume were limited in this study, speed data was available for almost all links. Therefore, speed was chosen as the base indicator in this study. Table 3.1 below also highlights the advantages of using speed as a metric, including its simplicity, stability, and ease of data acquisition.

Table 3.1: Congestion Metrics and Assessment Criterias (Rao and Rao, 2012)

	Simplicity	Ease of Data Collection	Stability	Repeatability	Management of Congestion	City Comparision	Continuous value
Speed	**O**	**O**	**O**	**O**	**O**	**X**	**O**
Travel Time	O	O	O	O	X	X	O
Delay	X	X	X	O	X	X	O
LOS and Volume	O	O	X	O	X	X	O

This study used the speed reduction index(SRI) as the basis for various metrics, focusing on network congestion. The following Equation 3.1 defines the index:

$$\text{Speed Reduction Index} = \frac{\text{85 th percentile speed} - \text{current speed}}{\text{85 th percentile speed}} \quad (3.1)$$

The SRI can have different values for different links, even at the same speed. As shown in Figure 3.4, if the original speed limit of a road is higher, the SRI will have a larger value. For example, when the speed on a left link in Figure 3.4 is measured as 40 km/h, the SRI will be 50% because the link's 85th percentile speed is 80 km/h. However, on the road with an 85th percentile speed of 50 km/h, 45 km/h corresponds to 20% on the SRI.

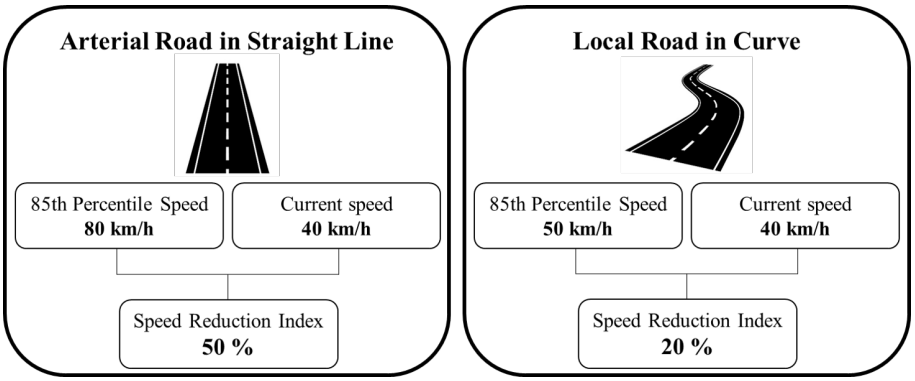


Figure 3.4: Example of SRI for different link

Using SRI, it is possible to measure how congested each road is while considering the relative hierarchy of the roads. Typically, the SRI is represented as the inverse of the speed graph. In the following examples, we will examine the characteristics of SRI.

During the morning and afternoon commute, congestion is worse at one time than the other. The links highlighted in red on the map on the right correspond to the blue link on the left graph, which is the source of the speed reduction causing the speed reduction on the downstream links. As shown in Figure 3.5 below, it can be seen that one of the SRI peaks is more severe than the other during the morning and afternoon commute.

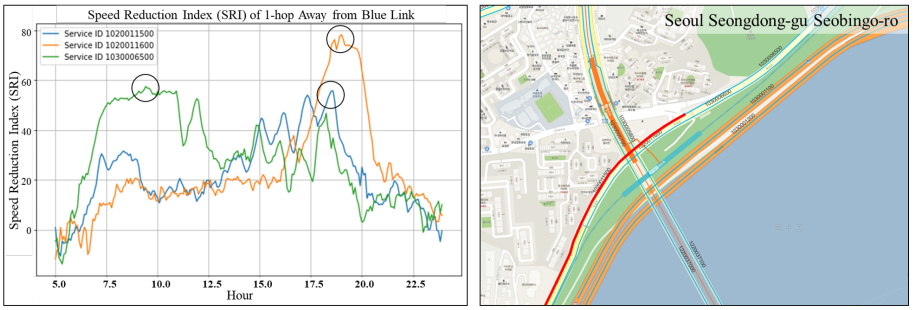


Figure 3.5: SRI peaks with different values in the morning and afternoon

The next thing we can infer is that not all connected links are necessarily related to congestion. Congestion propagation may not occur for links that involve U-turns or right/left turns. The figure below, Figure 3.6, illustrates the speed reduction propagation originating from the link indicated by the blue graph on the left. It can be seen that even when the SRI value rises, the yellow graph remains at a low value. This is because there are different types of connectivity and connectivity that have no relationship to each other.

In Figure 3.6, again, we can confirm that speed reduction propagation takes longer than expected. Shockwave propagation is different from speed reduction propagation. The literature shows that the speed of an urban shockwave is 13.32 km/h, but the speed observed in our graph is much slower than that (Ramezani and Geroliminis, 2015). This means shockwave and speed reduction is a different phenomenon.

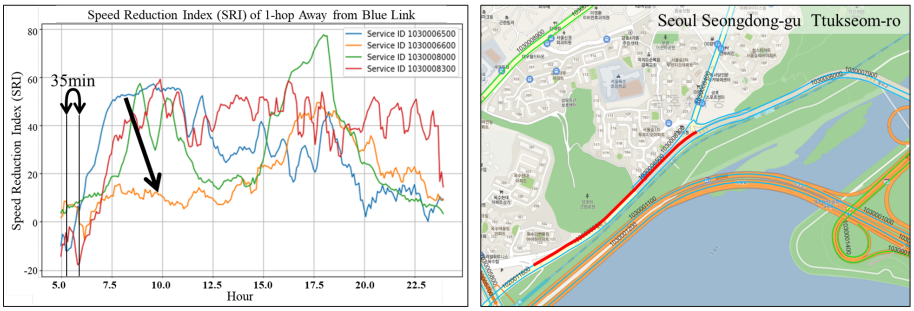


Figure 3.6: Discovery of unrelated links and delay in speed reduction propagation

Lastly, it can be difficult to intuitively understand speed reduction propagation beyond 1-hop, especially when considering various possible routes for speed reduction propagation. This is due to the interconnectedness of the network, which can make it difficult to discern a clear trend in congestion propagation. This highlights the need for more sophisticated modeling techniques to

accurately predict and analyze speed reduction propagation in complex urban networks. By developing more advanced models, we can gain a better understanding of how speed reduction propagates and identify effective strategies for mitigating its effects.

3.1.3 Creating an Impact on Adjacent Network Index

The impact on the adjacent network index aims to assess whether the current speed reduction on a given link results in future speed reduction on connected links. This index is calculated as the product of the current speed reduction index of the given link and the sum of the future speed reduction index of the links connected to the given link.

The current speed reduction index of the given link plays a crucial role in this calculation, as an uncongested road cannot cause congestion on other links. This consideration ensures that only congested links are evaluated when determining the impact on the adjacent network, thus providing a more accurate representation of the potential traffic issue.

The future speed reduction index of connected links represents the extent of the future speed reduction in the neighborhood due to the speed reduction of the target link. By accounting for the speed reduction index on connected links, the index offers a comprehensive understanding of how the current congestion on a specific link may contribute to the overall traffic conditions in the surrounding area.

There was an issue regarding how many hops to consider in the index. Realistically, it is challenging to assume that speed reduction propagates consistently beyond three hops on urban roads. Even the SRI graph becomes uncertain in terms of correlation after passing just two hops. If congestion were to propagate beyond three hops, the propagation pattern would ultimately be reflected

in the indices, as we will observe the speed reduction for all links up to three hops away. The relationship between a specific link and a link five hops away can be determined through the impact on the adjacent network index for the two hops away link.

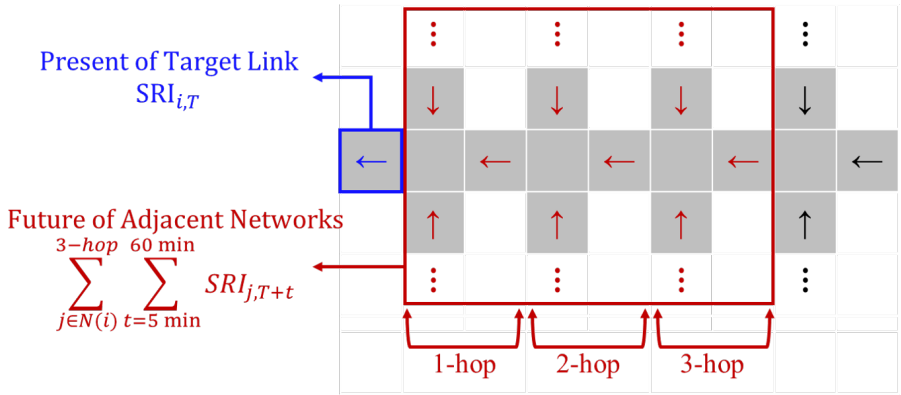


Figure 3.7: Diagram of impact on adjacency matrix index

In summary, taking into account more than three hops would likely not provide additional valuable information, as the relationship between links becomes less clear and less directly connected with each hop. Focusing on the immediate neighboring links (up to two or three hops away), the impact on the adjacent network index can provide more accurate and actionable insights into speed reduction and its propagation through the road network. As a result, the index has been set up as Equation (3.2) below.

Impact on Adjacent Network Index, of Link i, at time T

$$= IANI_{i,T} = \sum_{j\in N(i)}^{3-hop} \sum_{t=5}^{60} SRI_{j,T+t} \quad (3.2)$$

3.1.4 Preprocessing the Index

Data engineering might be just as important, if not more so, than model development. The data distribution was adjusted to ensure that it was suitable for smooth model training. The initial distribution of SRI is shown on the left side of Figure 3.8. Most values are between -20 and 100; however, some values have large negative numbers. Some values even reach -575. These values occur because some vehicles drive anomalously fast on roads with significantly low 85th-percentile speeds. Our primary focus is on speed reduction, not determining the speeds of fast-moving vehicles. Therefore, we performed normalization by appropriately reducing the absolute values of such data points.

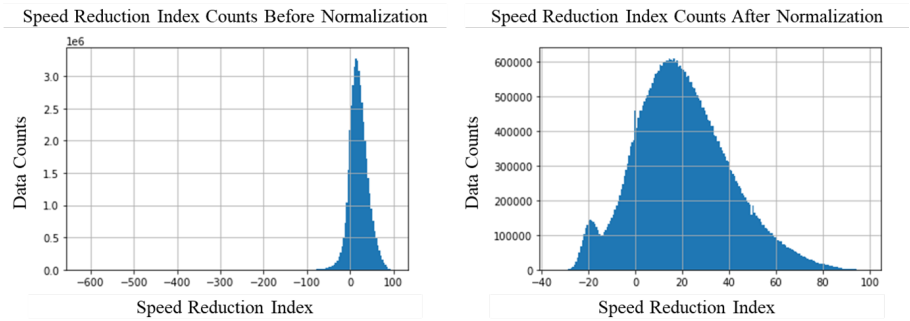


Figure 3.8: Before and after normalizing SRI

The previously validated SRI serves as input data for the model. Next, the distribution of data we examined is the model's output, which is the sum of future SRI values in the adjacent network. As seen on the left side of Figure 3.9, since the SRI was processed relatively well, the sum of future SRI values exhibited a distribution close to normal. However, by applying a square root transformation, we were able to make the distribution even more closely resemble a normal distribution.

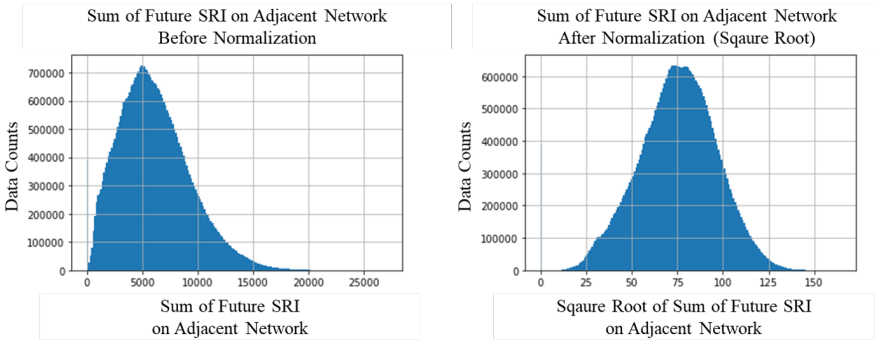


Figure 3.9: The Sum of future SRI before and after normalization

3.2 Analyzing the IANI

3.2.1 Statistical Property of IANI

Table 3.2 represents the pre-normalization values of the sum of future SRI for 5,068 links over 28 days, with 24 hours per day and 12 values per hour, resulting in 40,868,352 data points.

Table 3.2: Statistics of a sum of future SRI

Statistics	Value
Mean	6,076.127
Standard Deviation	3,219.975
Min Value	0.000
25-Percentile Value	1.061
50-Percentile Value	3,720.780
75-Percentile Value	5,733.657
Max Value	27,155.070

The average value is 6,076.127, with a standard deviation of 3,219.975. The

preprocessing step ensures that the minimum value of SRI is set to 0, resulting in 0 values in the sum of future SRI. The median is 3,720.780, and the maximum value is 27,155.070. The IANI calculation involves applying depreciation coefficients of 0.925 and 0.8 for the time step and number of hops, respectively. The statistics for each hop of the links are provided below.

Table 3.3: Statistics about the number of N-hop links

Statistics	1-hop	2-hop	3-hop	Sum 1 to 3-hop
Mean	2.798	7.213	15.074	26.085
Standard Deviation	1.232	3.647	7.355	11.699
Min Value	0.000	0.000	0.000	1.000
25-Percentile Value	2.000	5.000	10.000	18.000
50-Percentile Value	3.000	7.000	15.000	26.000
75-Percentile Value	4.000	10.000	20.000	34.000
Max Value	8.000	23.000	42.000	66.000

On average, each link receives contributions from approximately 26.1 links. IANI is calculated by multiplying the sum of future SRI by the individual link's SRI. The histogram in Figure 3.10 illustrates the distribution of IANI using SRI and normalized SRI.

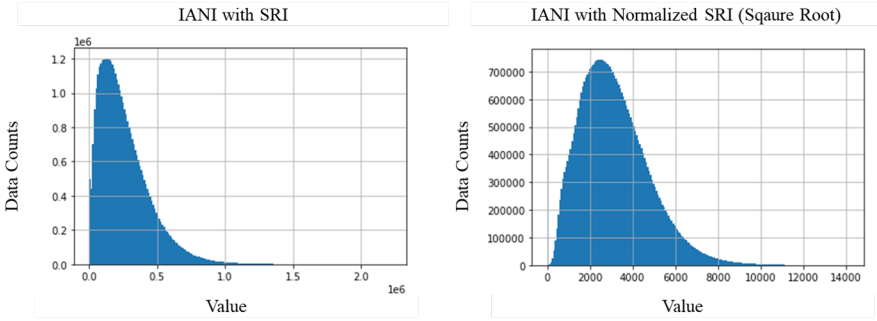


Figure 3.10: IANI with SRI and normalized SRI

The detailed statistics of IANI values are presented in Table 3.4. It can be observed that the distribution is left-skewed for the un-normalized case, indicating a higher concentration of values towards the lower end.

Table 3.4: Statistics of IANI with SRI and normalized SRI

Statistics	IANI with SRI	IANI with Normalized SRI
Mean	258,779.8	3,121.9
Standard Deviation	189,006.0	1,626.1
Min Value	0.0	0.0
25-Percentile Value	118,976.9	1,929.8
50-Percentile Value	217,374.9	2,919.7
75-Percentile Value	353,851.6	4,088.9
Max Value	2,225,545	14,150.1

When visualizing the IANI values, Figure 3.11 is obtained. It shows higher values in areas corresponding to the central business district, such as near Gangnam and Yeouido. Arterial links such as Dongbu Expressway also showed a high value of IANI.

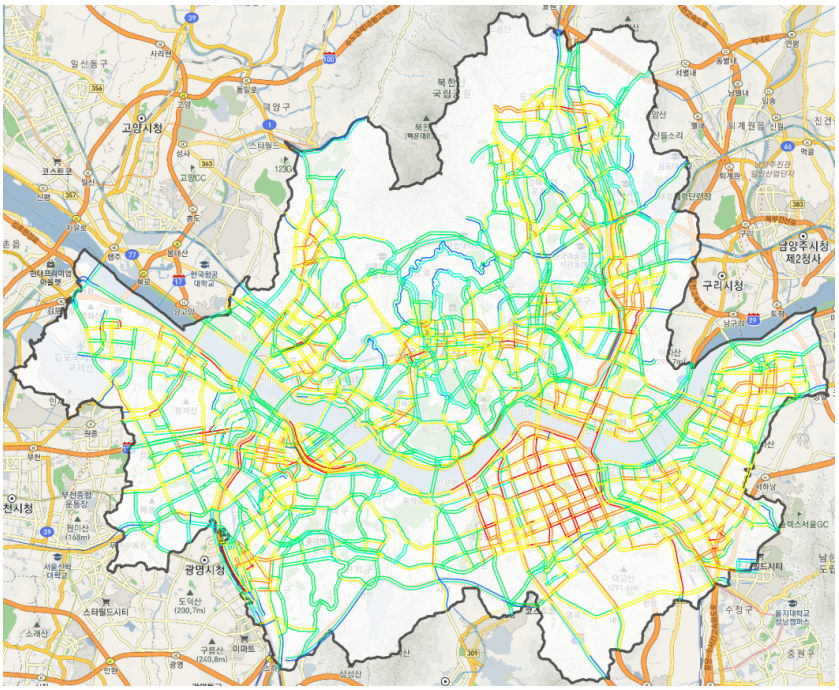


Figure 3.11: Visualizing the IANI value

The top 10 roads with high IANI values are as in Table 3.5. Notably, many of these roads are major arterial roads, including the Gangbyeon Expressway, Gyeongbu Expressway, and Dongbu Expressway, along with regular roads like National Assembly Road.

The difference between averaging IANI without distinguishing the time and day of the week and averaging it separately for morning and afternoon rush hours and weekdays and weekends is significant. The following contents in Table 3.6 are the top 10 links during weekdays in the morning hours (6-10 AM). It shows a stronger concentration on major arterial roads.

Table 3.7 illustrates the top 10 links during weekdays in the afternoon hours (5-9 PM). It can be observed that the top 10 links are more diverse, encompassing various roads.

Table 3.5: Roads with the highest IANI mean value

Rank	Name of the Road	Speed Limit (km/h)	Length (m)	Number of Lanes
1	Dongbu Expressway (동부간선도로)	57.5	1588.7	2.2
2	Dongbu Expressway (동부간선도로)	80.0	369.3	2.0
3	Gangbyeon Expressway (강변북로)	80.0	2892.8	4.0
4	Gyeongbu Expressway (경부고속도로)	70.0	1698.9	4.0
5	Banpo-daero (반포대로)	50.0	618.2	2.5
6	Olympic Expressway (올림픽대로)	80.0	1742.3	4.6
7	National Assembly-daero (국회대로)	50.0	1188.6	2.2
8	Dongbu Expressway (동부간선도로)	60.0	310.4	2.0
9	Dongbu Expressway (동부간선도로)	60.0	200.6	1.2
10	Dongbu Expressway (동부간선도로)	80.0	2964.3	3.0

Table 3.6: Roads with the highest IANI mean value at weekday and morning(6-10am) peak

Rank	Name of the Road	Speed Limit (km/h)	Length (m)	Number of Lanes
1	Dongbu Expressway (동부간선도로)	57.5	1588.7	2.2
2	Dongbu Expressway (동부간선도로)	80.0	4459.1	3.0
3	Olympic Expressway (올림픽대로)	80.0	3796.9	4.2
4	Dongbu Expressway (동부간선도로)	80.0	369.3	2.0
5	Dongbu Expressway (동부간선도로)	80.0	2964.3	3.0
6	Dongbu Expressway (동부간선도로)	80.0	2553.0	3.0
7	Dongil-ro (성동구 동일로)	52.0	1352.6	2.9
8	Gangbyeon Expressway (강변북로)	80.0	2892.8	4.0
9	Gyeongbu Expressway (경부고속도로)	70.0	1698.9	4.0
10	Dongbu Expressway (동부간선도로)	80.0	639.2	3.0

Table 3.7: Roads with the highest IANI mean value at weekday and afternoon(5-9pm) peak

Rank	Name of the Road	Speed Limit (km/h)	Length (m)	Number of Lanes
1	Dongbu Expressway (동부간선도로)	57.5	1588.7	2.2
2	Seocho-daero (서초구 서초대로)	50.0	568.1	3.0
3	Gyeongbu Expressway (경부고속도로)	70.0	1698.9	4.0
4	National Assembly-daero (영등포구 국회대로)	73.3	2216.5	2.4
5	National Assembly-daero (영등포구 국회대로)	50.0	1188.6	2.2
6	Olympic Expressway (올림픽대로)	80.0	1742.3	4.6
7	Banpo-daero (서초구 반포대로)	50.0	618.2	2.5
8	Dongbu Expressway (동부간선도로)	80.0	3332.4	3.0
9	Dongil-ro (성동구 동일로)	50.0	605.5	2.5
10	Bongeunsa-ro (강남구 봉은사로)	50.0	624.3	3.0

Table 3.8 shows the top 10 links during weekends in the morning hours (6-10 AM). Due to relatively lower traffic volume compared to other time periods, peripheral roads are more prominent in the top links. This aligns with the results obtained from past speed prediction, where the attention values of peripheral roads in the outskirts of Seoul were found to be higher during weekends.

The contents of Table 3.9 are the top 10 links during weekends in the afternoon hours (5-9 PM). Once again, it is evident that major arterial roads are selected as important links.

Based on the findings, it can be observed that different links are important in different time periods. There are two main strategies that can be considered based on this index.

The first strategy is to expand the links with high IANI value from a transportation planning perspective. By physically widening the links, it is expected that congestion caused by limited space can be alleviated.

The second strategy is to implement vehicle route diversion. Avoiding routes with high predicted future IANI values makes it possible to prevent the influx of traffic exceeding the road capacity. However, the formulation of precise strategies based on signal utilization is difficult since the data does not include turn-type information.

Table 3.8: Roads with the highest IANI mean value at weekend and morning(6-10am) peak

Rank	Name of the Road	Speed Limit (km/h)	Length (m)	Number of Lanes
1	Seosomun-ro (중구 서소문로)	50.0	198.8	2.0
2	Gosanja-ro (동대문구 고산자로)	50.0	406.3	3.0
3	Myeongil-ro (강동구 명일로)	40.0	276.0	1.0
4	Geumnanghwa-ro (강서구 금낭화로)	50.0	149.4	2.0
5	Songi-ro (송파구 송이로)	30.0	300.8	2.0
6	Dongbu Expressway (동부간선도로)	80.0	369.3	2.0
7	Seobu Expressway (서부간선도로)	80.0	1109.4	1.8
8	Hangeulbiseok-ro (노원구 한글비석로)	30.0	488.1	1.5
9	Dongnam-ro (강동구 동남로)	50.0	327.7	3.0
10	Seooreung-ro (은평구 서오릉로)	50.0	125.7	3.0

Table 3.9: Roads with the highest IANI mean value at weekend and evening(5-9pm) peak

Rank	Name of the Road	Speed Limit (km/h)	Length (m)	Number of Lanes
1	Dongbu Expressway (동부간선도로)	57.5	1588.7	2.2
2	Gangbyeon Expressway (강변북로)	80.0	2892.8	4.0
3	Olympic Expressway (올림픽대로)	80.0	1742.3	4.6
4	Gyeongbu Expressway (경부고속도로)	70.0	1698.9	4.0
5	Dongbu Expressway (동부간선도로)	80.0	2964.3	3.0
6	Banpo-daero (서초구 반포대로)	50.0	618.2	2.5
7	Dongbu Expressway (동부간선도로)	80.0	369.3	2.0
8	Dongil-ro (성동구 동일로)	50.0	605.5	2.5
9	National Assembly-daero (영등포구 국회대로)	50.0	1188.6	2.2
10	Olympic Expressway (올림픽대로)	80.0	1515.2	5.0

3.2.2 Comparing IANI with Graph Centralities

The section explores the relationship between graph centralities and IANI. Each centrality has the following characteristics: Degree centrality represents how directly connected a link is to other links. Katz centrality indicates how many different paths can reach other links. Closeness centrality measures how close the distance is from other links to the target link. Betweenness centrality determines whether a specific link is part of the shortest path between two other links.

After examining the correlation coefficients between various centralities and IANI, it was found that all coefficients were positive. A strong relationship with IANI was observed between degree centrality and Katz centrality, which are closely related to the direct connection index with other links.

Table 3.10: Correlation between various centralities and IANI

	Betweenness Centrality	Degree Centrality	Katz Centrality	Closeness Centrality	IANI
Betweenness Centrality	1.000	0.126	0.087	0.348	0.120
Degree Centrality		1.000	0.837	0.180	0.630
Katz Centrality			1.000	0.201	0.573
Closeness Centrality				1.000	0.196
IANI					1.000

However, further analysis of links with high IANI values revealed additional features in addition to these characteristics. As we can check in Table 3.10, it is observed that betweenness centrality is characteristically high for links with extensive value of IANI. This is likely due to the specific properties of each centrality measure. Betweenness centrality assesses whether a road is included in the shortest path between two other roads.

The fact that a road greatly impacts its surrounding roads implies that it has a high probability of being included in the shortest path between other roads, which is an intuitive notion. However, it is difficult to determine whether or not other centrality measures should be incorporated, as the reasons for their inclusion or exclusion are not clear. Further analysis and investigation of these centrality measures may provide valuable insights into their relevance and potential contributions to the overall understanding of the road network.

Table 3.11: Value of various centralities of the roads with the top 100 INAI

Type of Centrality	Top 100 mean	Mean	Ratio
Betweenness Centrality	0.0114	0.0031	3.7249
Degree Centrality	0.0020	0.0015	1.3109
Katz Centrality	0.0099	0.0129	1.6016
Closeness Centrality	0.0684	0.0617	1.1094

3.2.3 Comparing IANII with SRI

The study compared IANI with commonly studied metrics in traffic speed analysis, SRI. Similar to the analysis conducted in Chapter 4.4.2, the research examined the correlation coefficients between each centrality and the metrics. The results are presented in Table 3.12, which shows that all network centralities exhibited higher correlation coefficients with IANI compared to SRI. This indicates that IANI incorporates not only the speed information but also the structural characteristics of the network.

Next, we grouped the values of SRI and IANI and examined the link attributes, such as length and number of lanes, within each group. Both SRI and IANI were divided into five quintiles. Figure 3.12 illustrates the variation of link attribute values by quintile, while Figure 3.13 shows the centralities' values for each quintile. The same information is summarized in Table 3.13 and 3.14.

Table 3.12: Comparing the correlation of SRI and IANI with various centralities

	Betweenness Centrality	Degree Centrality	Katz Centrality	Closeness Centrality	SRI	IANI
Betweenness Centrality	1.000	0.126	0.090	0.348	0.031	0.120
Degree Centrality		1.000	0.837	0.180	0.018	0.630
Katz Centrality			1.000	0.201	0.020	0.573
Closeness Centrality				1.000	0.008	0.196
SRI					1.000	0.378
IANI						1.000

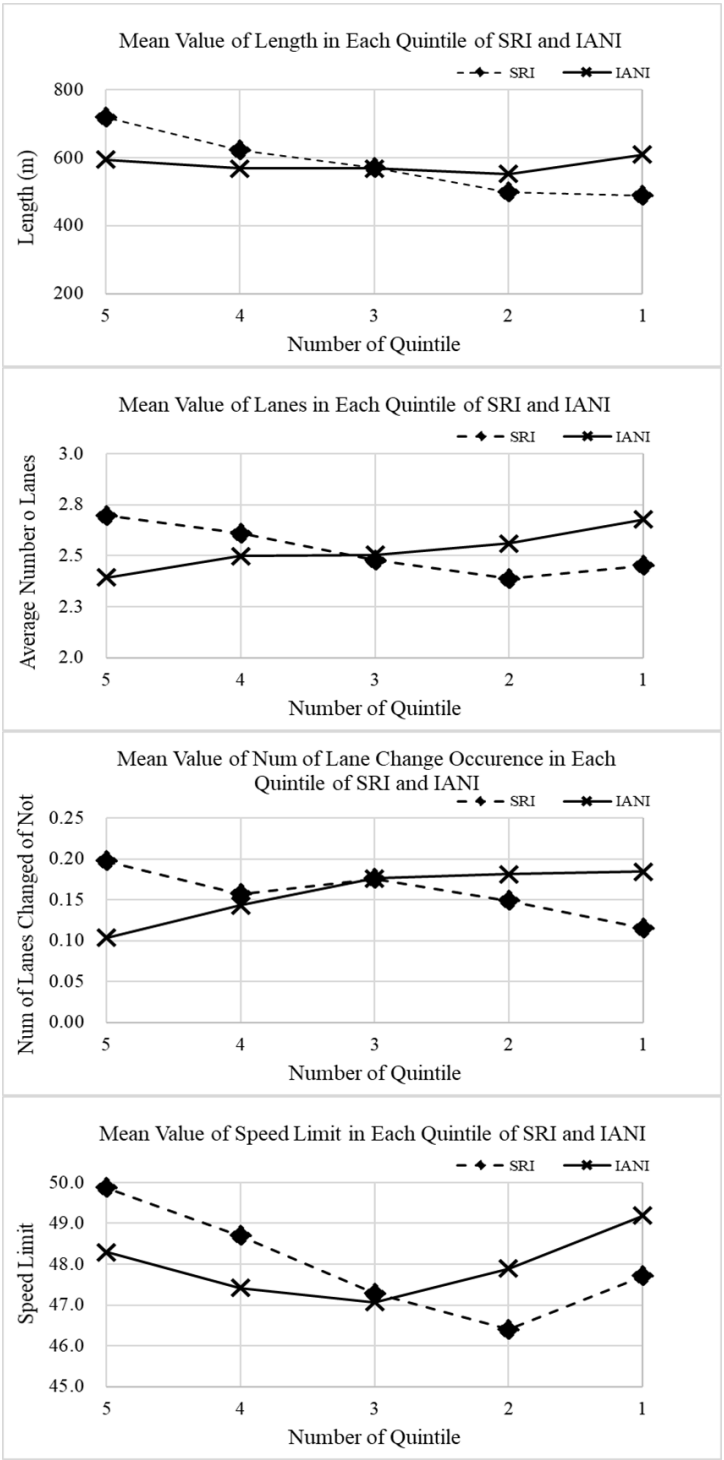


Figure 3.12: Mean values of various criteria by quintile of SRI and IANI

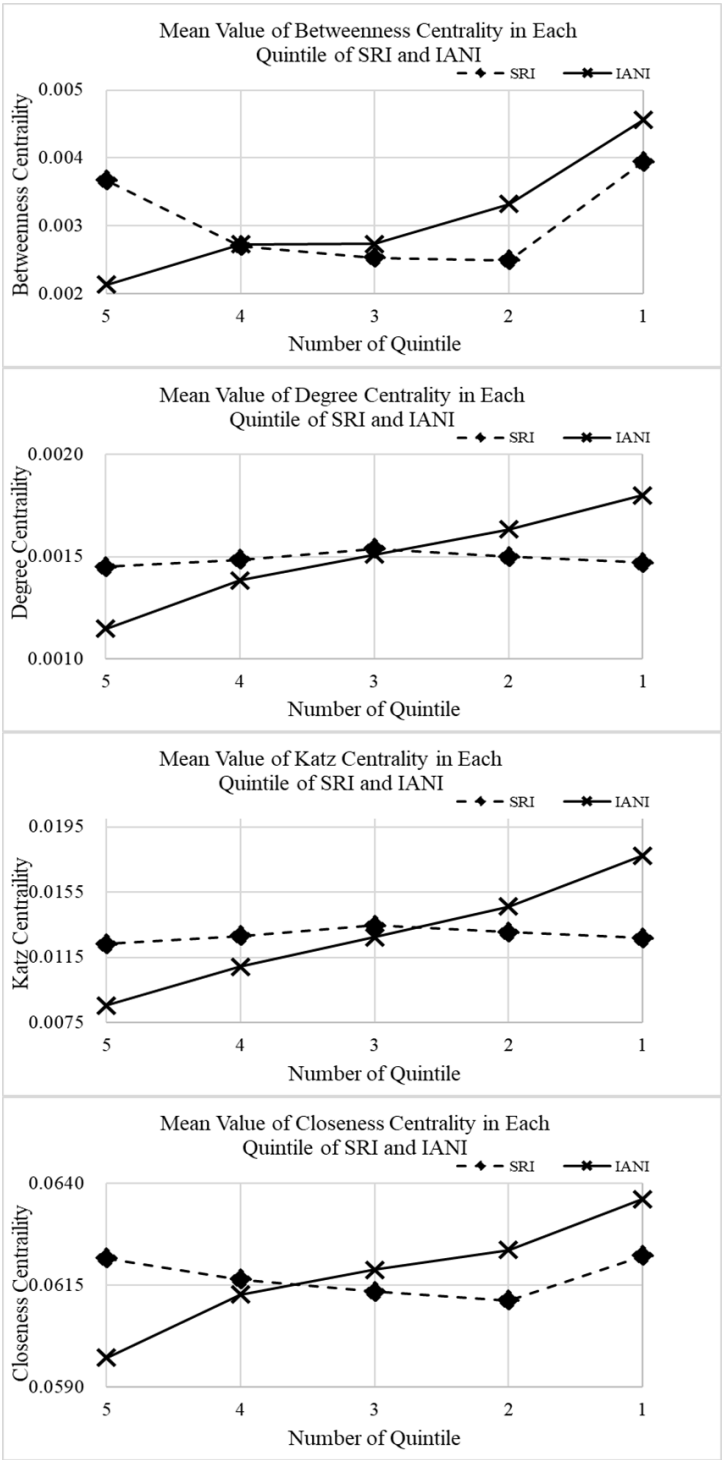


Figure 3.13: Mean values of various centralities by quintile of SRI and IANI

Table 3.13: Various value of links by quintile of IANI

Quintile	1	2	3	4	5
Length	488.01	499.31	569.36	622.31	718.48
Average Number of Lanes	2.451	2.387	2.479	2.609	2.697
Lane Chaged or Not	0.115	0.149	0.176	0.157	0.197
SRI	48.375	43.230	40.581	38.029	34.439
IANI	3600.2	3292.1	3172.8	2931.4	2612.8
Speed Limit	47.711	46.393	47.288	48.695	49.869
Betweenness Centrality	0.0039	0.0025	0.0025	0.0027	0.0037
Degree Centrality	0.0015	0.0015	0.0015	0.0015	0.0015
Katz Centrality	0.0127	0.0130	0.0135	0.0128	0.0123
Closeness Centrality	0.0622	0.0611	0.0613	0.0616	0.0622

For IANI, it was observed that as the quintile increased, the link length, the number of lanes, and the occurrence of lane changes all increased. Conversely, SRI exhibited the opposite trend, indicating that narrower roads experienced more localized speed reductions. Various centralities consistently increased across IANI quintiles, indicating that links with higher IANI values are more strategically positioned within the network.

Just like the previous section categorized the value into quintile groups, categorical values are generally commonly used in transportation planning and operation. However, when utilizing continuous values, it is possible to examine the magnitude of extreme values while also dividing the values into categories using appropriate thresholds. This approach allows us to take advantage of both the benefits of continuous values and the categorization aspect.

Figure 3.14 compares cases where SRI and IANI exhibit different patterns in

Table 3.14: Various value of links by quintile of SRI

Quintile	1	2	3	4	5
Length	609.33	552.22	567.76	568.36	594.22
Average Number of Lanes	2.677	2.56	2.505	2.499	2.394
Lane Chaged or Not	0.184	0.181	0.176	0.143	0.104
SRI	44.62	41.404	40.284	39.377	38.809
IANI	4395	3613	3165.4	2694	1891.3
Speed Limit	49.196	47.895	47.071	47.422	48.293
Betweenness Centrality	0.0046	0.0033	0.0027	0.0027	0.0021
Degree Centrality	0.0018	0.0016	0.0015	0.0014	0.0011
Katz Centrality	0.0177	0.0146	0.0127	0.0109	0.0086
Closeness Centrality	0.0636	0.0624	0.0619	0.0613	0.0597

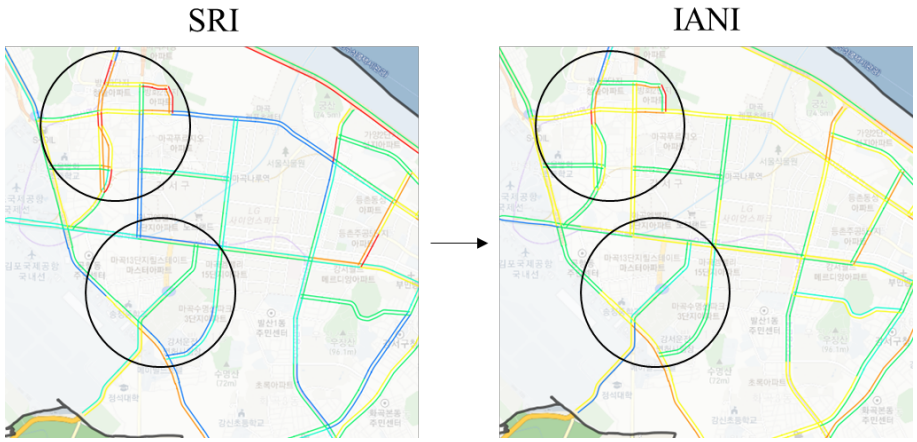


Figure 3.14: The different values between SRI and IANI in the same region

the same region. For SRI, it tends to be higher for shorter and narrower roads, suggesting higher values assigned to areas with limited connectivity, such as apartment access roads. On the other hand, major roads leading to arterial

routes tend to have lower SRI values. In contrast, IANI considers both speed reduction and network importance, thus selecting less important roads and more significant roads from a network perspective.

Chapter 4

Graph Attention Model for Urban Network

4.1 Background of the Graph Attention Model

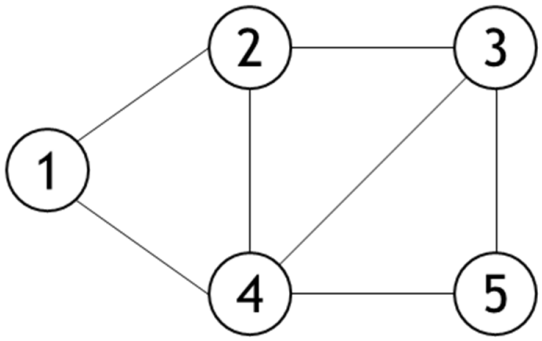
In the prediction problem, the attention mechanism selects data to be referred to with a higher weight using the higher attention value. The attention mechanism consists of a query, key, function, and value.

$$A(q, K, V) = \sum_i \text{softmax}(f(q, K))V \tag{4.1}$$

A query is input data of a target we are trying to predicate. In this paper, the current link speed data or a speed reduction index becomes a query. The relationship between a query and several other keys is determined by a predefined function, where a key is a non-query link. Function determines the relationship between a query and a key. In GAT, a function is a 1-layer Neural Network (NN).

$$A(q, K, V) = \sum_i \text{softmax}(1 - \text{ layer Nerual Network } (q \| K))V \tag{4.2}$$

The detailed process can be illustrated using the following example, which depicts a graph composed of five nodes and seven edges.



Adjacency Matrix

node	1	2	3	4	5
1	0	1	0	1	0
2	1	0	1	1	0
3	0	1	0	1	1
4	1	1	1	0	1
5	0	0	1	1	0

Figure 4.1: Example graph with five nodes and seven edges

Let the feature vector in layer l of any node i be $h_i^{(l-1)}$. Here, the relationship between the feature vector $h_i^{(l-1)}$ of node i and the feature vector $h_j^{(l-1)}$ of node j is defined as the following Equation (4.3) in the term energy e. The attention value α is the softmax of this energy value for the connected node. Equation (4.4) shows the softmax process in the GAT model.

Attention values are obtained through a simple softmax function. Matrix W is utilized to obtain energy, and since W is a learnable parameter, higher performance energy is achieved as the iteration progresses. The 1-layer NN of Equation (4.3) below is also a learnable parameter, with a total of two matrices being learned in the process of obtaining energy. The same matrix is used for all node pairs. Therefore, the existing matrix W can be used continuously even if a new node is added. The graph neural network is explained to have inductive

properties. The opposite is called the transductive property. The processes of Equation 4.3) and Equation (4.4) below are performed only for adjacent nodes. Since we are currently checking node i, Equation (4.3) and (4.4) are performed for $k \in N(i)$, the nodes included in the neighbor of node i.

$$e_{ij}^{(l-1)} = \text{LeakyReLU}\left(1 - \text{layer}\, NN\left(\boldsymbol{W}h_i^{(l-1)} \| \boldsymbol{W}h_j^{(l-1)}\right)\right) \tag{4.3}$$

$$\alpha_{ij}^{(l-1)} = \frac{\exp\left(e_{ij}^{(l-1)}\right)}{\sum_{k\in\mathcal{N}(i)} \exp\left(e_{ik}^{(l-1)}\right)} \tag{4.4}$$

The attention value α, obtained through Equation (4.4) is used to calculate the feature vector $h_i^{(l)}$ of time step i. In this case, the activation function is applied after adding the feature vectors to the neighboring nodes. In the original paper, where graph attention networks were introduced, Leaky Relu was used, and in this study, the convention of the original paper was followed (Velickovic *et al.*, 2017).

$$h_i^{(l)} = \text{ ActivationFunction } \left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(l-1)} \boldsymbol{W}h_i^{(l-1)}\right) \tag{4.5}$$

Figure 4.2 illustrates the process of applying the graph attention mechanism to node 2. Node 2 is connected to nodes 1, 3, and 4, so the energy for these nodes is obtained by passing their feature vectors through matrix W and feeding the concatenated result through the 1-layer Feedforward Neural Network. Using this energy, the attention value α is calculated. By performing a weighted sum based on α, the feature vector $h_2^{(l)}$ of node 2 of layer l is obtained. In a practical implementation, this involves a linear combination of the feature vectors $h_2^{(l-1)}$ from all previous layers $(l-1)$.

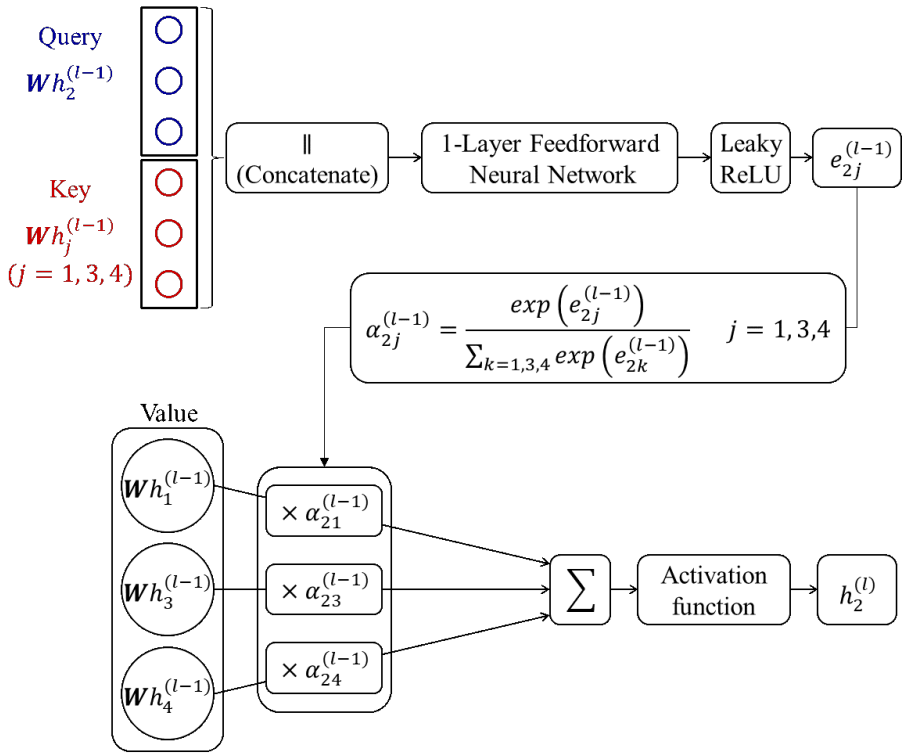


Figure 4.2: Calculating the feature vector of node 2 in layer l

The above explanation discusses the original graph attention mechanism introduced by Velickovic et al. (2017) (Velickovic *et al.*, 2017). This paper's model is based on an improved model called Attention-Based Spatial-Temporal Graph Convolutional Networks (ASTGCN). ASTGCN is a modified graph attention model designed explicitly for spatiotemporal prediction (Guo *et al.*, 2019). ASTGCN applies spatial attention to the same time step and temporal attention to different time steps. The fundamental structure of ASTGCN is similar to the graph attention network, except that it seeks attention along two axes. As shown in Figure 4.3, temporal attention is applied to individual nodes, while spatial attention is applied to a single time step.

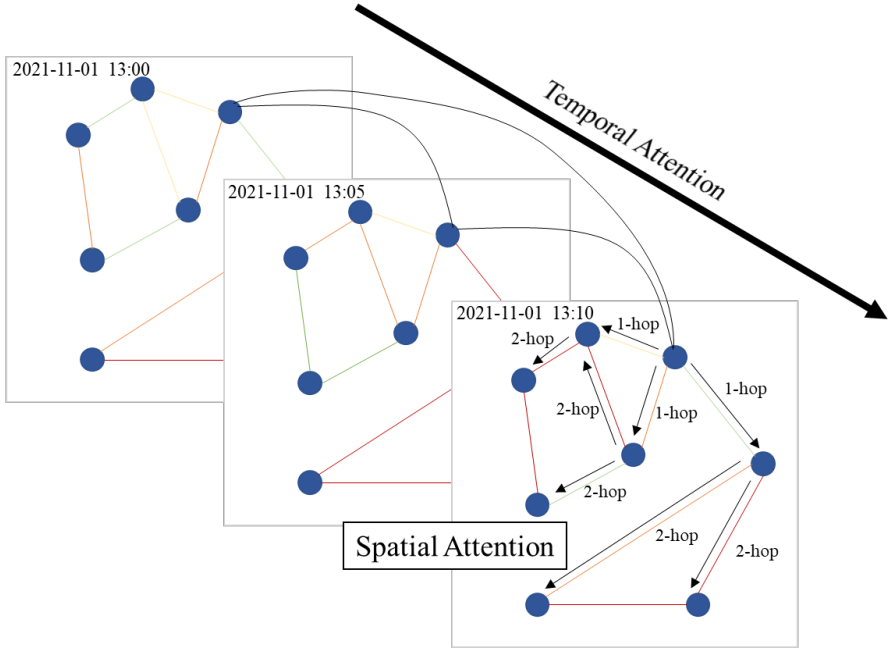


Figure 4.3: Temporal and spatial attention of ASTGCN

The core structure of ASTGCN consists of connecting multiple spatiotemporal attention blocks. In the original ASTGCN paper, separate spatiotemporal attention blocks were constructed for time, date, and day of the week. However, this study uses a single block to prevent overfitting. Additionally, the skip connection structure is omitted.

Spatial attention is denoted as S, while temporal attention is denoted as T. The multiplications of matrices depicted below can be understood as a single-layer perceptron.

$$\mathbf{S} = \mathbf{E} \cdot \sigma \left(\left(\mathbf{h}^{(1-1)} \mathbf{W}_1 \right) \mathbf{W}_2 \left(\mathbf{h}^{(1-1)} \mathbf{W}_3 \right)^{\mathrm{T}} + \mathbf{b}^{\mathrm{l}} \right) \tag{4.6}$$

$$\boldsymbol{S}'_{i,j} = \frac{\exp\left(\boldsymbol{s}_{i,j}\right)}{\sum_{k \in N(i)} \exp\left(\boldsymbol{s}_{i,k}\right)} \tag{4.7}$$

$$\mathbf{T} = \mathbf{F} \cdot \sigma \left(\left(\mathbf{h}^{(1-1)} \mathbf{U}_1 \right) \mathbf{U}_2 \left(\mathbf{h}^{(1-1)} \mathbf{U} \right)^T + \mathbf{c}^1 \right) \tag{4.8}$$

$$\boldsymbol{T}'_{i,j} = \frac{\exp\left(\boldsymbol{T}_{i,j}\right)}{\sum_{k \in N(i)} \exp\left(\boldsymbol{T}_{i,k}\right)} \tag{4.9}$$

ASTGCN's attention is computed using matrix multiplication, as described above. Matrix W and matrix U serve as learnable spatial and temporal attention parameters, respectively. While the process is similar to graph attention networks, an additional matrix multiplication step is included for dimensional unification.

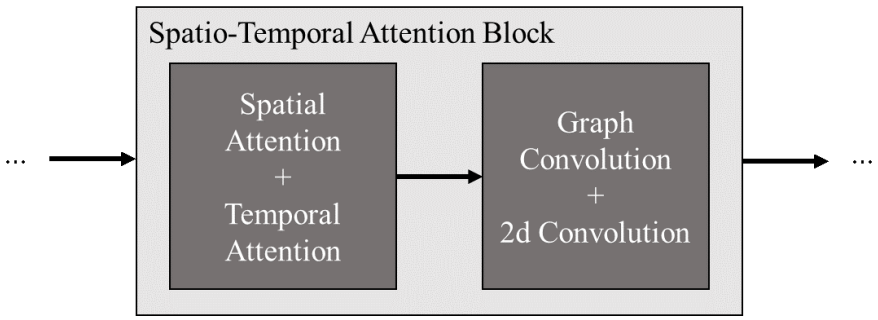


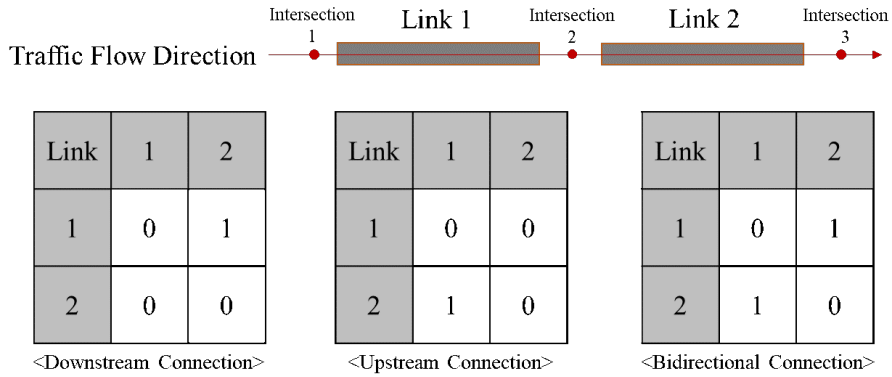
Figure 4.4: Iterative scheme of Spatio-temporal attention block

After obtaining spatial and temporal attention through Equations (4.6), (4.7), (4.8), and (4.9), the spatiotemporal attention block is completed by connecting with the graph convolution and 2D convolution layers. As the dimension of the feature vector entering and leaving these blocks remains constant, an appropriate number of blocks can be added based on the complexity of the problem. In this study, the research was conducted using two blocks for developing the adjacency matrix and a single block for developing the prediction model.

4.2 Improving the Graph Attention Model with Adjacency Matrix

4.2.1 The Traffic Flow Awareness Adjacency Matrix

The issue of the Euclidean distance-based adjacency matrix not accurately recognizing traffic flow was addressed by introducing a traffic flow-aware adjacency matrix. As shown in Figure 4.5, the traffic flow direction begins at Intersection 1, proceeds through Link 1, enters Intersection 2, and then goes through Link 2 and Intersection 3.



Link	1	2
1	0	1
2	0	0

<Downstream Connection>

Link	1	2
1	0	0
2	1	0

<Upstream Connection>

Link	1	2
1	0	1
2	1	0

<Bidirectional Connection>

Figure 4.5: The traffic flow awareness adjacency matrix

In the situation described above, three connection matrices have been proposed: the Downstream connection matrix, the Upstream connection matrix, and the Bidirectional connection matrix. In the Downstream connection matrix, Link 2 is situated downstream of Link 1 in the direction of traffic flow. Thus, in this case, since link 2 corresponds to the Downstream connection of link 1, a value of 1 is assigned to the (1, 2) position in the matrix. All other values remain at 0. The Upstream connection matrix operates similarly. Given that link 2 is located downstream of link 1, a value of 1 is assigned to the (2,

1) position in the matrix. The Bidirectional connection matrix combines both Downstream and Upstream connections.

There are also papers that have conducted research using adjacency matrices, such as the Bidirectional connection matrix, in existing traffic state prediction studies (Li *et al.*, 2017). However, this paper is unique in differentiating Downstream and Upstream connections. The underlying concept stems from the notion that the amount of influence a specific link receives from other links may differ. Even in the case of interrupted flow due to urban links, if there are no additional congestion factors besides traffic signals, the speed of the upstream section will propagate to the downstream section as-is, making a Downstream connection more appropriate. Conversely, if traffic congestion in the downstream section is severe and the traffic decrease in the downstream section continues to the upstream section as a form of shockwave, applying an Upstream connection would be more suitable. Thus the selection of an adjacency matrix may differ by the type of problem.

4.2.2 Introducing Katz Centrality to the Adjacency Matrix

Katz centrality is a measure representing the centrality of a node applied in graph theory (Katz, 1953). It represents the sum of all possible walk lengths of a specific node. Katz centrality mainly indicates the relative power of a node's influence on others in a social network. It can be expressed in a formula, as shown in Equation (4.10).

$$C_{\text{Katz}}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \beta^k (A)_{ji}^k \tag{4.10}$$

The above expression has the same form as the sum of the geometric series of a matrix. β is a decay function that prevents the values of the above ex-

pression from diverging. If an appropriate β cannot be selected, the value itself may diverge, and computation becomes impossible. Since it is the sum of the geometric series, assuming that k goes to positive infinity, it can be arranged in a closed form, as shown in Equation (4.11) below. If Equation (4.10) diverges, the inverse of Equation (4.11) cannot be obtained, resulting in an incalculable expression. However, if k is not infinite, Katz centrality can be obtained unconditionally.

$$C_{Katz}(i) = \left(I - \beta A^T\right)^{-1} - I \tag{4.11}$$

The idea behind Katz centrality is both intuitive and powerful. It is evident that nearby links have a more substantial impact on one another. Therefore, this study designed an adjacency matrix based on Katz centrality. At this time, the number of hops of the link to be connected is determined according to the value of k. Determining an appropriate number of hops has a significant impact on the performance of the model. An explanation of this is illustrated in Figure 4.6.

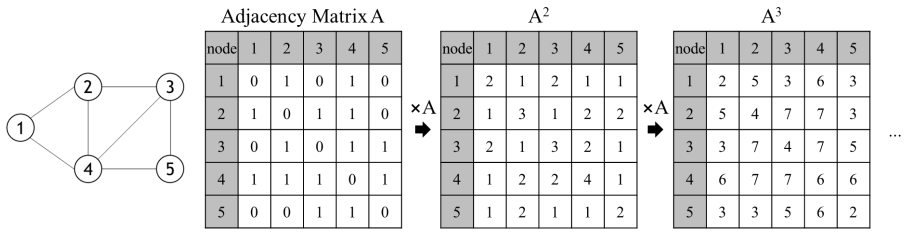


Figure 4.6: k-times multiplying adjacency matrix enables to connect k-hop matrix

The adjacency matrix for the graph on the left in Figure 4.6. is shown on the right. Node 1 and Node 2 have a value of 1 because they are connected to each other, and Node 1 and Node 3 have a value of 0 because they are not connected.

However, if you multiply A once to make A^2, we can see that there is number 2 in the (1, 3) position that describes the connectivity of Node 1 and Node 3. This is the number of cases that can lead to a 2-hop connection. There is one path to Node 1 – Node 2 – Node 3 and another path through Node 1 – Node 4 – Node 3. In this way, A^2 guides the way to a 2-hop connection. A^k represents the number of cases that can go to a k-hop connection. A^k inevitably diverges if a decay hyperparameter such as β is not added. Therefore, normalization should be performed together with the decay hyperparameter.

Another advantage of introducing Katz centrality is that the number of graph convolution layers can be reduced. When defining the adjacency matrix as A matrix, the k-layer is required to express the k-hop connection, but when $\sum_{k=1}^{\infty} \beta A^k$ is introduced, it can be reached in 1-layer. Reducing the number of layers also helps prevent overfitting. When a large number of layers are introduced to investigate a simple phenomenon, the number of parameters is excessive, and overfitting occurs. As explained in Section 3.3.4, Limitations in Graph Attention Model, simple phenomena such as traffic are even more vulnerable to overfitting. In that respect, it can be said that it is more appropriate to introduce the concept of Katz centrality to this research problem. For additional information on this, see Section 4.3.3, Handling the Overfitting and Oversmoothing Problem.

Section 3.3.4, Limitations in Graph Attention model, also pointed out the high regularity of the road network. Because of this, the average path length increases. If Katz centrality is applied, a k-hop connection can be connected in one layer to reflect a wide range of networks.

4.2.3 Handling the Overfitting and Oversmoothing Problem

The overfitting problem mentioned in Section 4.3.2., Introducing Katz Centrality to the Adjacency is directly related to the number of parameters. Therefore, this study solved the problem by reducing the spatiotemporal attention block to two or one. In addition to this, a small number of epochs were applied.

In addition to the overfitting problem, graph neural networks have another issue: the oversmoothing problem (Chen *et al.*, 2020). The oversmoothing problem refers to an issue in which the values of all nodes become the same when the receptive field of the graph neural network becomes too wide (Liu *et al.*, 2020). The phenomenon is shown in Figure 4.7. Each time it passes through one graph convolutional layer, it can reflect connections that are 1-hop further away. However, if too many connections are expressed compared to the graph, the receptive field of all nodes becomes almost the entire graph.

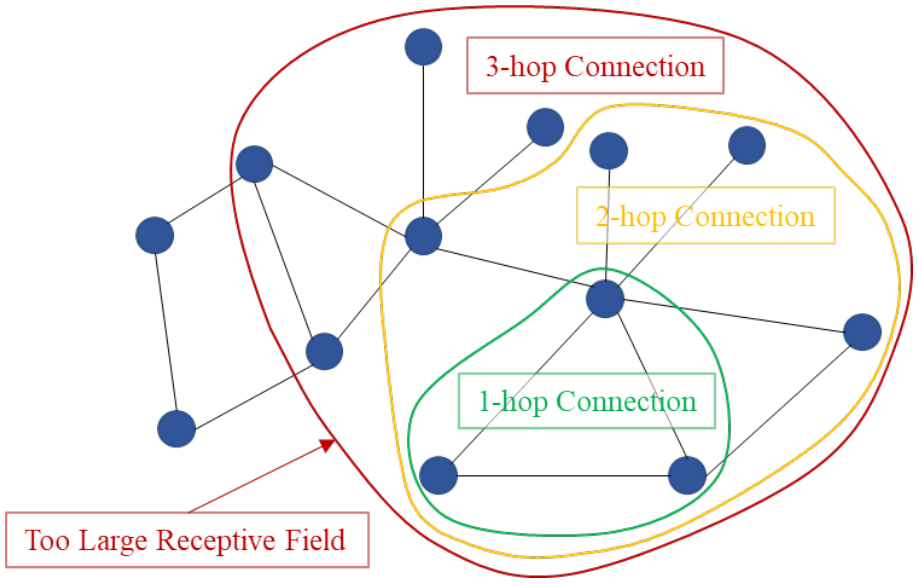


Figure 4.7: The example of oversmoothing caused by extensive receptive field

Graph neural networks operate using message-passing techniques to share information with neighboring nodes. When a message is received from all nodes, a problem occurs where the nodes' features in the graph become very similar. To prevent this issue, a small number of layers should be maintained. The road network is highly regular and has a longer average path length than other graphs, making it less susceptible to oversmoothing. However, even so, oversmoothing can occur at the sub-network level. For this reason, the number of layers in the entire network was kept small.

4.3 Adding Physical Meaning to the Model

4.3.1 Reflecting the Heterogeneity of Road Networks

The first physical meaning employed in this study is to reflect the road network's heterogeneity. Interrupted and uninterrupted flow, which makes the hierarchy of road networks, shows different aspects of speed reduction propagation.

Uninterrupted flow refers to links without signals and those that are straight, stretching continuously. Consequently, the traffic speed on these roads is relatively fast, and the relationships between the roads are comparatively straightforward. When a speed reduction occurs downstream, it is directly transmitted upstream. The propagation of speed reduction moves faster compared with the interrupted flow. Figure 4.8 below illustrates the Eastern Expressway in the Seongdong District. The colors blue, green, yellow, orange, and red indicate increasing levels of speed reduction, respectively. Traffic proceeds from the top to the bottom of the figure. In each row, as time passes, it is evident that congestion is distinctly propagating backward.

Figure 4.9 illustrates the case of interrupted flow in Gangnam. In the case of interrupted flow, it is difficult to determine a clear direction of congestion

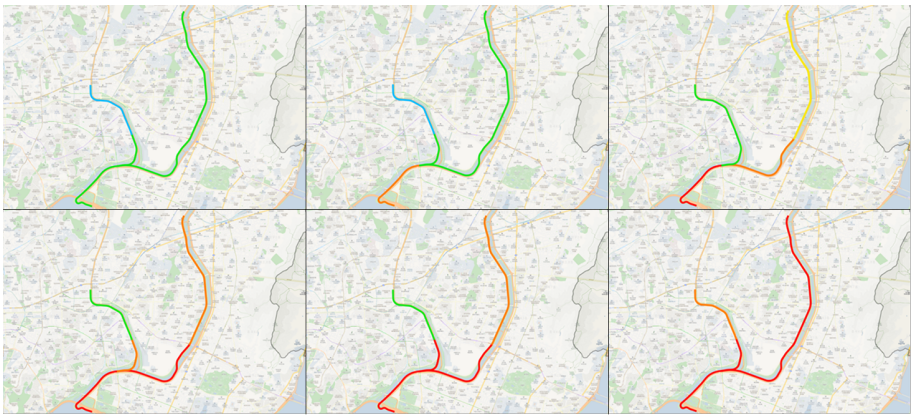


Figure 4.8: Propagation of speed reduction through an uninterrupted flow. Each subfigure's time step starts from 6:00 am to 8:00 30 min interval

propagation. The road network is densely interconnected, making it challenging to identify the cause and origin of congestion. As time passes, speed reduction continually interacts between connected roads, making it virtually impossible to pinpoint the source of congestion.

Upon analyzing the 5,068 service links in Seoul, a total of 280 uninterrupted flow links and 4,788 interrupted links were identified. The classification results are shown in Figure 4.10 below. As can be seen in the figure, the distinction between the major expressways within the urban center and the remaining roads is clearly visible, highlighting the importance of considering both types of flows when modeling traffic patterns.

In order to reflect the distinct characteristics of these different flows, this study proposes the Attention-based Spatio-Temporal Heterogeneous Graph Convolution Network (AST-HGCN). This novel approach takes into account the heterogeneous nature of urban traffic flows, capturing the unique relationships and dynamics present in interrupted and uninterrupted flows. By incorporating these considerations into the model, the AST-HGCN provides a more accu-

Figure 4.9: Propagation of speed reduction through an interrupted flow. Each subfigure's time step starts from 6:00 am to 8:00 30 min interval

rate and comprehensive understanding of traffic patterns, ultimately resulting in improved performance and more precise predictions in a variety of traffic scenarios.

The model structure is shown in Figure 4.11 below. An ASTGCN block containing information for all links was created, along with separate ASTGCN blocks for uninterrupted flow and interrupted flow information. From the block containing the total link data, the output vector h is obtained; from the block containing the uninterrupted flow link data, the output vector u is obtained; and finally, from the block containing the interrupted flow link data, the output vector i is obtained.

The output vector h obtained from the total link data was divided into vector slices corresponding to uninterrupted and interrupted flows. Let us call these vectors h_i and h_u for interrupted flow and uninterrupted flow, respectively. We then obtained the attention between h_i and i, as well as the attention between

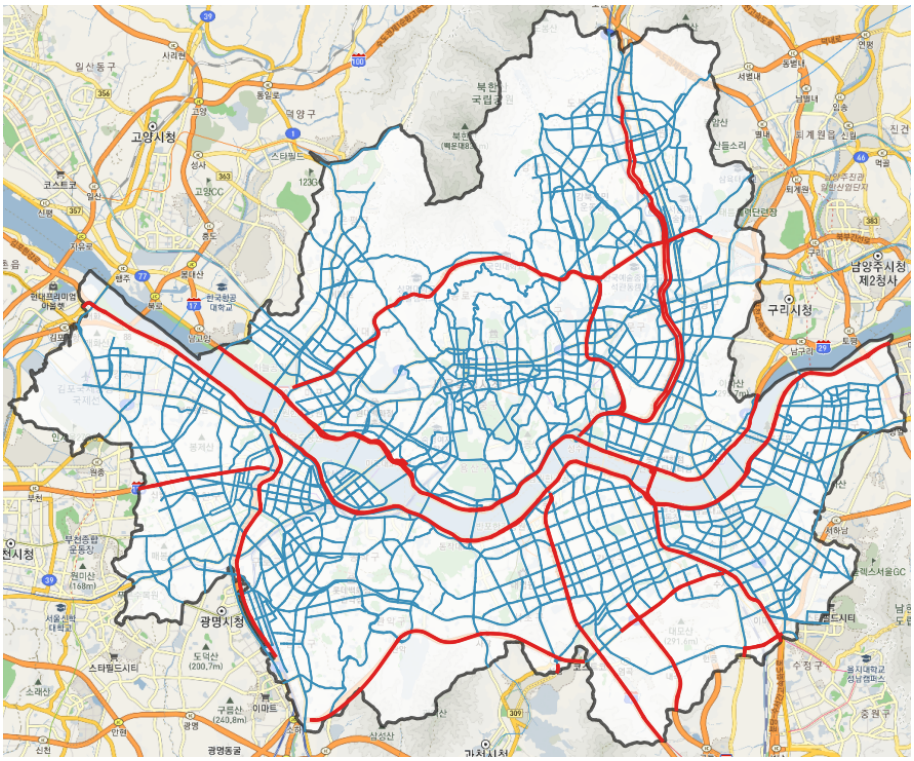


Figure 4.10: Visualization of uninterrupted flow(red) and interrupted flow(light blue)

h_u and u. Subsequently, after receiving the output vector from each block, attention values were obtained for each type of flow using a 1-layer feedforward neural network. Finally, as shown in Equation 4.12, each attention value was multiplied by the output vector, and a weighted sum was calculated. The weighted sum of each output vector becomes the output $\hat{y}$, and is compared with true y to obtain various loss values.

$$\hat{y} = h + \alpha_{u,\hbar_u} u + \alpha_{i,\hbar_i} i \tag{4.12}$$

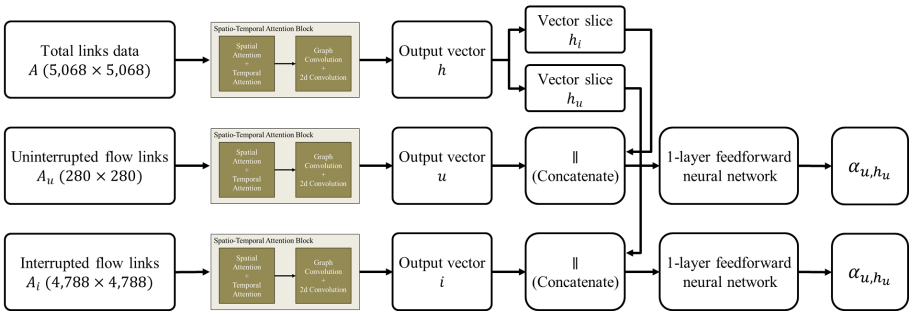


Figure 4.11: The structure of Attention-based Spatio-Temporal Heterogeneous Graph Convolution Network (AST-HGCN)

4.3.2 Incorporating Traffic Volume Data

In the current data acquisition location, the number of traffic volume sensors is significantly lower compared to the traffic speed. However, traditional speed prediction models commonly incorporate traffic volume. Therefore, this study aims to explore possibilities for integrating traffic volume into our model. Figure 4.12 displays the locations where traffic volume is collected, and it can be observed that there is a significant scarcity of traffic volume data compared to traffic speed data.

The scarcity of data can be addressed using deep learning techniques. In deep learning, data compression can be achieved using an autoencoder composed of an encoder and decoder. We assume that the traffic volume data has already been compressed into a hidden embedding, and we can utilize this embedding by feeding it into the decoder. The overall framework of the model is depicted in 4.13.

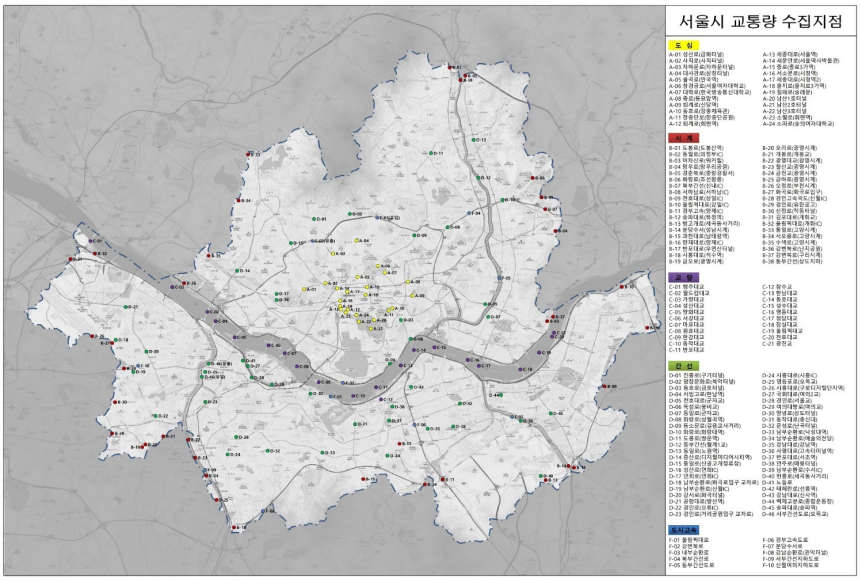


Figure 4.12: Data points with traffic volume

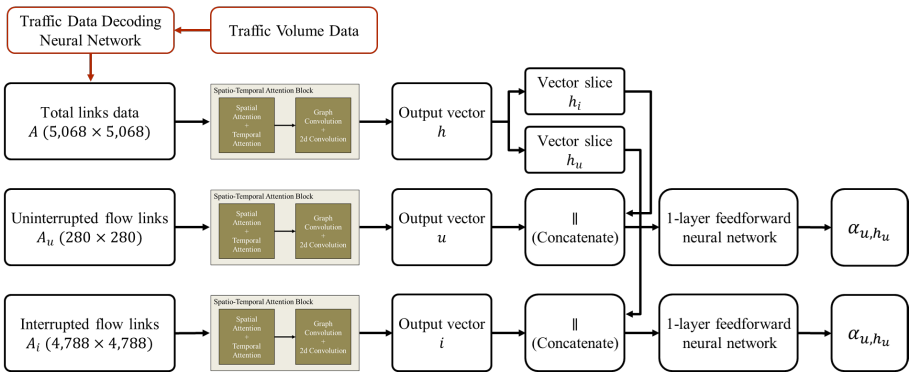


Figure 4.13: Incorporating traffic volume using a decoder neural network

4.3.3 Adding a Penalty as an Attention Loss

There are some issues concerning attention values. Attention values are parameters, and the primary objective of a model is to produce an output vector, not to fit parameters to give them a physical meaning. This can lead to biased results of attention value, as the focus is not on generating accurate attention values but on the model performance.

There is no significant issue when the number of reference links is small. In this study, the example of Gangnam-gu does not pose a significant problem. However, as the number of links increases, the individual links begin to get confused about which links they should reference. Moreover, each link's SRI trend is similar: high at the commute time and relatively low at the other time. So when the model face with the challenge of choosing a few links from among 5,000 for each link, the values do not converge easily.

In such cases, the attention values tend to concentrate on a single link. It is not just a slight concentration; all the links focus solely on that specific link. An example of this is shown in Table 4.1. Looking at Link 2 in Table 4.1, one can see that all the attention values are concentrated on it. The problem is that Link 2 is not actually an important link, and this phenomenon occurs randomly.

Additional loss terms can be introduced to address this problem and generate more realistic attention values. These loss terms include the concentration penalty and the distance penalty. The concentration penalty aims to prevent the attention value from being overly focused on a specific random link, ensuring a more balanced distribution of attention across links.

The distance penalty encourages the model to assign higher attention values to nearby links rather than distant ones. By incorporating these penalties, the model is guided to generate more realistic attention values that better represent

Table 4.1: Example of an attention value matrix that is concentrated on a specific link

	Link 1	**Link 2**	Link 3	**...**	Link n-1	Link n
Link 1	0.00017	**0.99812**	0.00011	...	0.00015	0.00011
Link 2	0.00021	**0.99141**	0.00027	...	0.00016	0.00023
⋮	⋮	⋮	⋮		⋮	⋮
Link n-1	0.00035	**0.99452**	0.00026	...	0.00018	0.00026
Link n	0.00017	**0.99275**	0.00018	...	0.00025	0.00019

the relationships between links in the road network.

$$\text{Concentration Penalty} = \sum_{j=1,2,\ldots,n} \left(\sum_{i=1,2,\ldots,n} \alpha_{i,j} \right)^2 \tag{4.13}$$

$$\text{Distance Penalty} = \sum_{\text{by row}} \boldsymbol{\alpha} \times \text{Shortest Path Matrix} \tag{4.14}$$

The total loss, including these penalties, is defined as follows:

$$\begin{aligned} \text{Loss} \ &= \text{RMSE loss} \ + \beta_1 \sum_{\text{flow type}} \text{Concentration Penalty} \\ &+ \beta_1 \sum_{\text{flow type}} \text{Concentration Penalty} \end{aligned} \tag{4.15}$$

Unfortunately, the model's performance inevitably decreases when this loss is introduced. This is because the model focuses less on the original objective since a different type of loss is included. Therefore, a decrease in prediction performance is unavoidable. We will later examine the implications that can be drawn from this model through the analysis of results.

Chapter 5

Results

5.1 Improving the Adjacency Matrix

This part of the research presents an improvement of the graph attention model using speed data from Gangnam-gu, which is the second step in the research workflow.

Before moving on to a more detailed discussion, the paper would like to introduce the overall prediction frame. The prediction frame in this study remains consistent throughout the experiments. Even if the prediction target changes to impact on adjacent network index from speed. Using the previous hour's data, the next hour's speed is predicted. The prediction forecasts the speed of all links within the network simultaneously. This approach is maintained even when dealing with future data. An example is shown in Figure 5.1. The first dataset consists of speed data input from 08:00 to 08:55, and the speed data from 09:00 to 09:55 is predicted. The second dataset uses speed data input from 09:00 to 09:55 and predicts speed data from 10:00 to 10:55. In this manner, a

result with a 1-hour input horizon and a 1-hour prediction horizon is obtained.

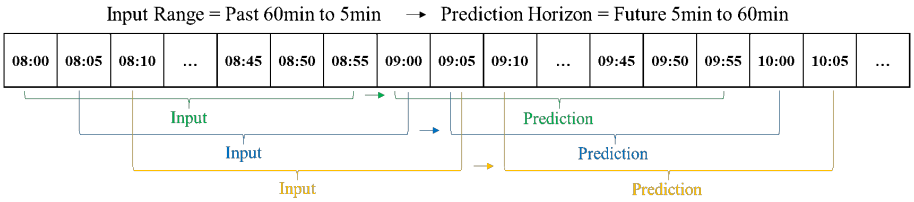


Figure 5.1: Input and prediction horizon and scheme

Thanks to this frame, a single prediction consists of a total of 12 data points, which is an hour. As shown ins Figure 5.2, data from future 5min to future 60min are predicted. The upper graph of Figure 5.2 displays the result of predicting the future 5min, and the graph below shows the result of predicting the future 60min. Predictions are depicted in red, while ground truth data are shown in blue. Overall, both future 5min and future 60min predictions exhibit appropriate performance. Qualitatively, it appears that the local peak of the future 5min is predicted more accurately than the local peak of the future 60min. Naturally, predicting the nearer future is an easier task. In fact, verifying whether the model better predicts the nearer future serves as another measure of whether the model has been properly trained.

Figure 5.3 displays the results. Over time, the RMSE loss exhibits a monotonically upward-sloping trend. This outcome aligns with our common understanding that we are better at predicting the near future and worse at predicting the far future. This graph shows that the model's speed prediction ability is an RMSE of 4.2 for the previous 5 min. This corresponds to a 13.64% MAPE error. Even when predicting the far future, the RMSE remains below 4.8 and demonstrates good performance.

From here on, the paper will discuss the results of the improved adjacency

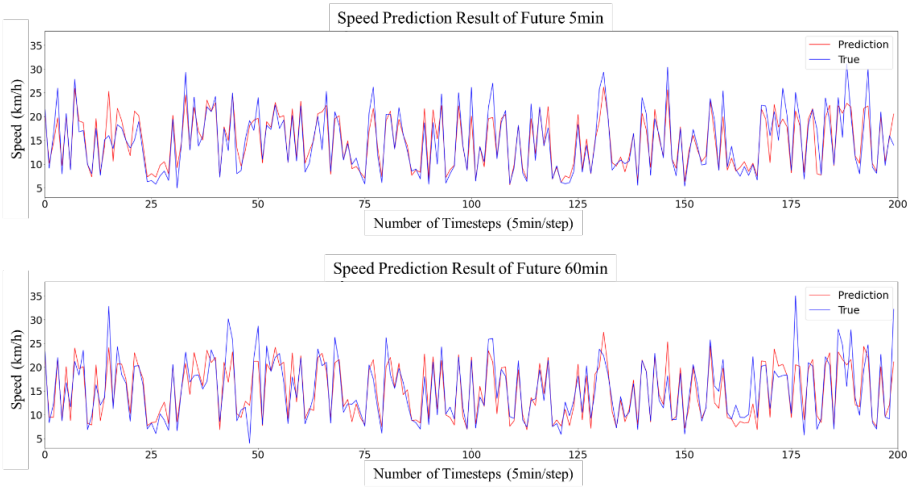


Figure 5.2: Speed prediction result of future 5min(top) and 60min(bottom) of Gangnam-gu link (selected)

matrix. By utilizing an adjacency matrix that combines the concepts of traffic flow awareness and Katz centrality, it demonstrated higher performance than the existing adjacency matrix. Detailed results can be found in Table 5.1.

To assess the accuracy and stability of the speed prediction data, 50 experiments were conducted for each type of connection. Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) were the error measurements used. In addition to evaluating the simple model performance, the standard deviation of the error was also checked to ensure the stability of model learning. The calculation process for RMSE and MAPE can be found in Equations 5.1 and 5.2.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_l})^2} \tag{5.1}$$

$$MAPE = \frac{100\%}{n}\sum_{\mathrm{i}=1}^{n}\left|\frac{y_{\mathrm{i}} - \widehat{y_{\mathrm{1}}}}{y_{\mathrm{i}}}\right| \tag{5.2}$$

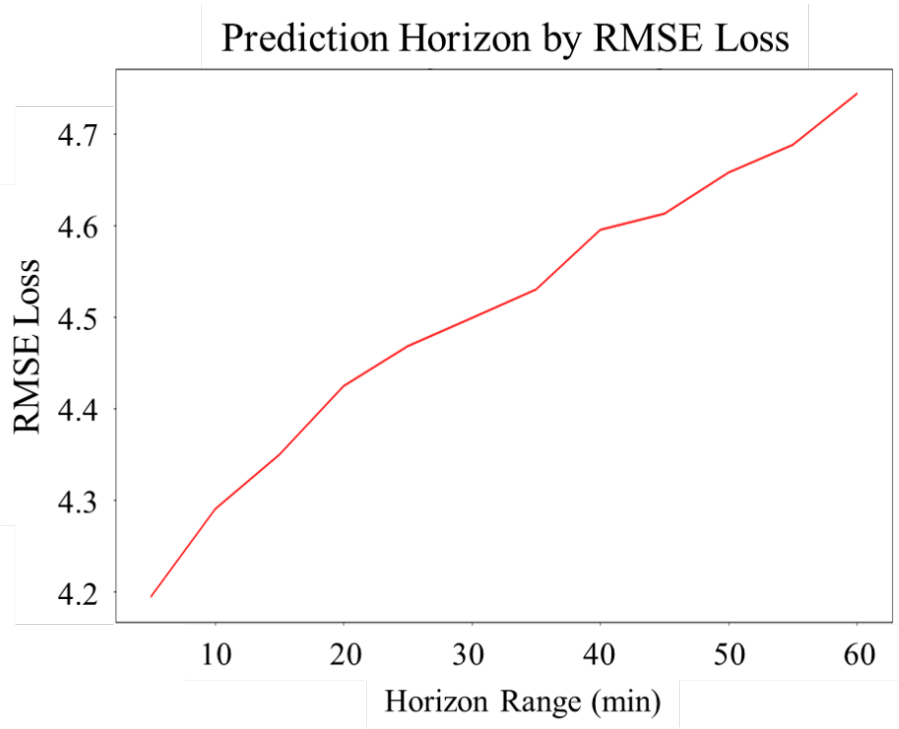


Figure 5.3: Prediction Horizon by RMSE loss of Gangnam-gu links

The first notable point in the results of Table 5.1 is that the 2D distance adjacency matrix based on Euclidean distance exhibited the lowest performance. When measuring speed with the Distance 2D adjacency matrix, the RMSE showed the worst performance, while the MAPE demonstrated the second-worst performance. This suggests that there is room for improvement in the model performance of numerous studies conducted using 2D distance so far. Surprisingly, the model's performance based on the 1D distance adjacency matrix was higher. This could be because the roads in Gangnam-gu are similar to a grid shape, and the 1D distance, which measures the distance along the side of the grid, accurately reflects the distance between links.

The Downstream 4-hop adjacency matrix displayed the best performance

Table 5.1: Speed prediction result in Gangnam-gu

Type of Connection	RMSE Avg.	RMSE Std.	MAPE Avg.	MAPE Std.
Downstream 1-hop	4.455	0.26	14.59	0.724
Downstream 2-hop	4.572	0.231	14.77	0.557
Downstream 3-hop	4.455	0.247	14.4	0.716
Downstream 4-hop	**4.437**	**0.239**	**14.41**	**0.565**
Upstream 1-hop	4.44	0.237	14.52	0.598
Upstream 2-hop	4.571	0.281	14.72	0.726
Upstream 3-hop	4.501	0.234	14.57	0.609
Upstream 4-hop	4.504	0.265	14.68	0.483
Bidirectional 1-hop	4.454	0.236	14.38	0.569
Bidirectional 2-hop	4.515	0.225	14.68	0.575
Bidirectional 3-hop	4.498	0.261	14.57	0.567
Bidirectional 4-hop	4.535	0.281	14.96	0.681
1D Distance	4.484	0.299	14.76	0.567
2D Distance	**4.634**	**0.265**	**14.82**	**0.684**

among the traffic flow-aware adjacency matrices. The Bidirectional adjacency matrix showed a trend of decreasing performance from 4-hop, which could be attributed to the effects of oversmoothing. This phenomenon was either absent or weakly apparent in the Downstream or Upstream adjacency matrix. As the Bidirectional adjacency matrix has the broadest receptive field, the probability of encountering oversmoothing issues is higher.

Before diving into further discussion, it is essential to clarify the meaning of the "attention value" that will be frequently mentioned from now on. The

attention mechanism introduces a weighted sum approach to the data prediction task. In the attention mechanism, the weight of the data deemed to have a more significant influence on the link through a specific function is increased. Consequently, the summation of the attention value referenced by a particular link is unconditionally 1. We cannot use an attention value greater than 1 in total. The left figure in Figure 5.4 shows an example of an attention value matrix obtained through learning. The center illustrates the weighted summation using attention value. At this point, the row summation is 1. We can determine the link's influence by summation on the column. It adds how many other links refer to the target link when predicting data for the next time step. The attention value claimed in this study is the result of this consensus.

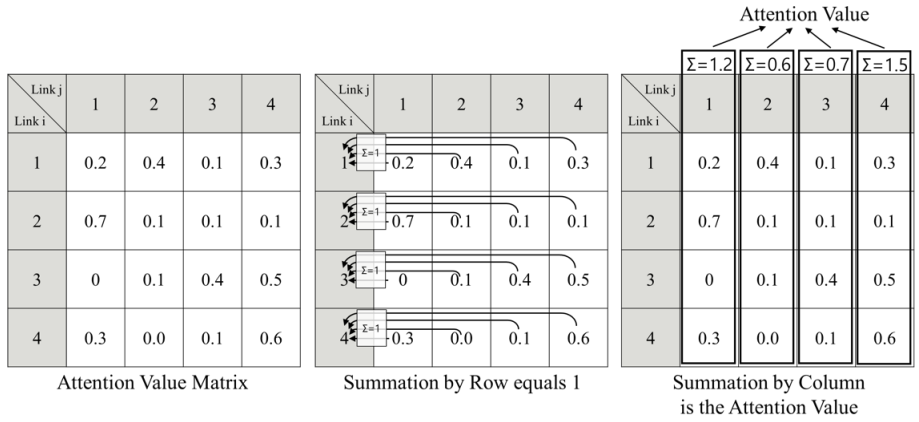


Figure 5.4: The illustration of the term "attention value"

Figure 5.5 presents the result of acquired attention values for each link targeting Gangnam-gu. The X-axis represents the link identification number with no physical meaning, while the Y-axis displays the summation of attention values. We can interpret the summation of attention values as the influence of each link. The link marked with 35 is the dominant figure, and based on this, we can

predict that this link is important. However, we should not be immediately persuaded by the absolute magnitude of the numbers. As mentioned several times before, attention values are very sensitive and can change significantly, even with minor adjustments. Therefore, it is more appropriate to observe changes in relative magnitude and attention value over time rather than attributing meaning to absolute values, which could lead to over-interpretation.

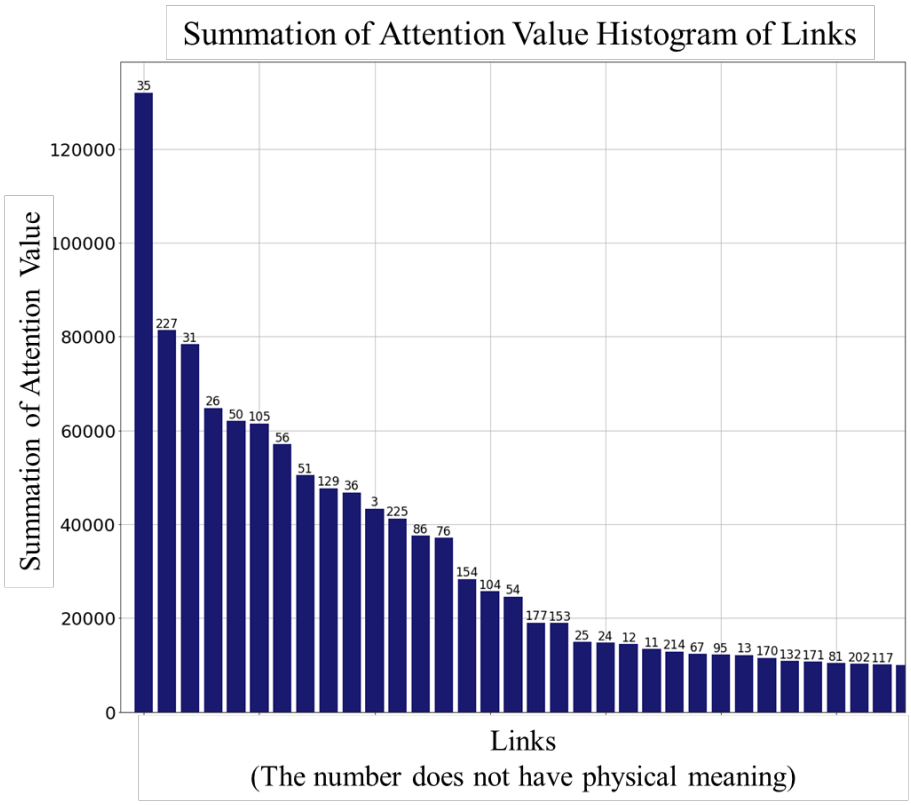


Figure 5.5: Attention Sum Histogram of Links in Gangnam-gu

Figure 5.6 shows the top 10 links of the summation of attention targeting Gangnam-gu. As the value of the summation of attention becomes similar as we go down to the lower level, it becomes difficult to conclude with certainty which

link is more important. However, we can judge that the above list's links are significant. One intuitive characteristic is that there are more significant links at the entry and exit points of the boundary than inside Gangnam-gu. Table 5.2 shows summary information for each road. For intuitive understanding, Korean notation is included in parentheses. The "Connecting region" column indicates the region to which the corresponding link is connected, and the "In/Out to Gangnam-gu" column indicates whether the corresponding link is a link entering or exiting Gangnam-gu. Links marked with "-" are within Gangnam-gu. In the case of "In/Out to Gangnam-gu," equal numbers were derived with four in-links and four out-links.

So, how should we interpret the summation of attention? A link with a large summation of attention value significantly impacts the road network. Due to the nature of downtown areas, roads with high traffic volume and congestion have a greater impact than roads with smooth traffic. Congestion mainly occurs during commuting hours, and the primary commuting route is located at the border between Gangnam-gu and other areas, not inside Gangnam-gu. For this reason, it is expected that priority links are mainly identified at the boundary between Gangnam-gu and other regions. For this hypothesis to be valid, it must be proven that the attention value at non-commuting hours is smaller than the attention value during the commute.

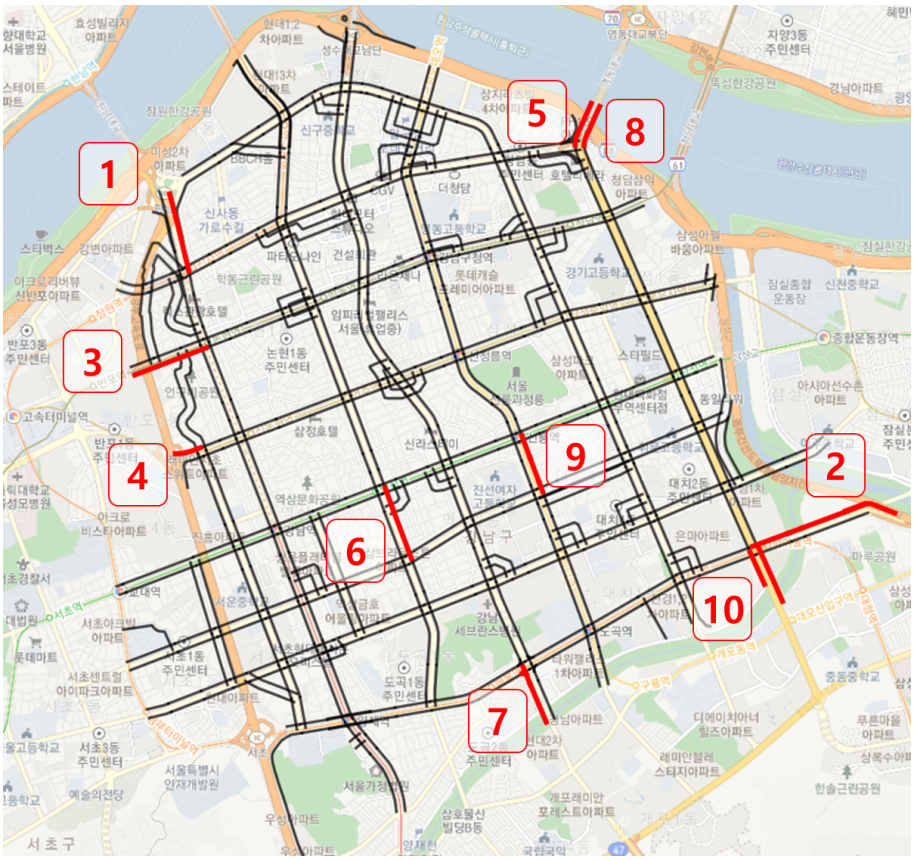


Figure 5.6: Top 10 links in Gangnam-gu by summation of attention

Table 5.2: Summary of Top 10 links in Gangnam-gu by summation of attention

Rank	Road Name	Connecting Region	In/Out to Gangnam-gu
1	Gangnam-daero (강남대로)	Shinsa-Hannam Bridge (신사-한남대교)	Out
2	Southern Beltway (남부순환로)	Songpa, Suseo-Daechi, Samsung (송파, 수서-대치, 삼성)	In
3	Sinbanpo-ro (신반포로)	Banpo-Nonhyeon (반포-논현)	In
4	Sapyeong-daero (사평대로)	Sinnonhyeon-Banpo IC (신논현-반포IC)	Out
5	Yeongdong-daero (영동대로)	Yeongdong Bridge - Cheongdam (영동대교-청담)	In
6	Nonhyeon-ro (논현로)	Maebong-Yeoksam (매봉-역삼)	-
7	Eonju-ro (언주로)	Dogok-Yangjae IC, Naegok IC (도곡-양재IC, 내곡IC)	Out
8	Yeongdong-daero (영동대로)	Cheongdam - Yeongdong Bridge (청담 -영동대교)	Out
9	Seolleung-ro (선릉로)	Hanti-Seongneung (한티-선릉)	-
10	Yeongdong-daero (영동대로)	Suseo IC-Samsung (수서IC-삼성)	In

Figure 5.7 displays the attention values obtained on Tuesday, November 2, 2021, for the top 3 links that showed the highest attention value. Attention values were obtained during the evening time when commuting occurs. In this figure, it can be observed that a consistently high attention value appears throughout the entire period. A more impressive insight can be obtained when comparing Figure 5.6 and Figure 5.7.

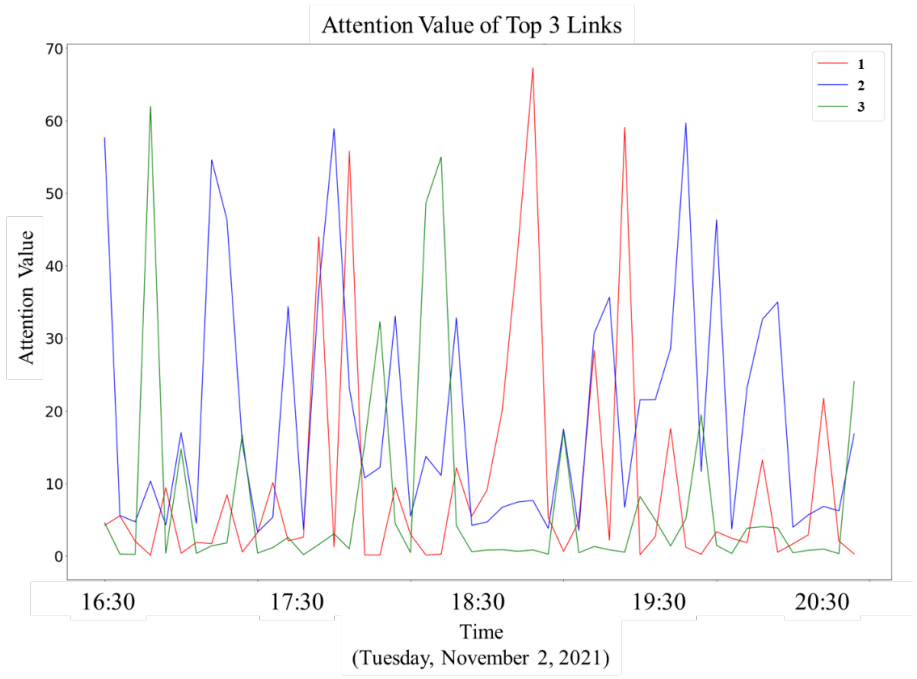


Figure 5.7: Attention value of top 3 links of Gangnam-gu on Tuesday, November 2, 2021, afternoon

Figure 5.8 presents the attention values of the top 3 links on Wednesday, November 3, 2021. Here, an "attention hole" with a low attention value of the top 3 links, which was not seen in Figure 5.7, is observed. This time corresponds to lunchtime during business hours, and it is a period when the demand for movement to the entrance and exit of Gangnam-gu is inevitably reduced. Most

of the traffic will be directed to restaurants or cafes, which can be reached by a walk. The occurrence of small attention values, or attention holes, on links entering Gangnam-gu during non-commuting hours, is consistent with our common sense. Although it is imperfect, this case study provides evidence that attention values propose relatively realistic values.

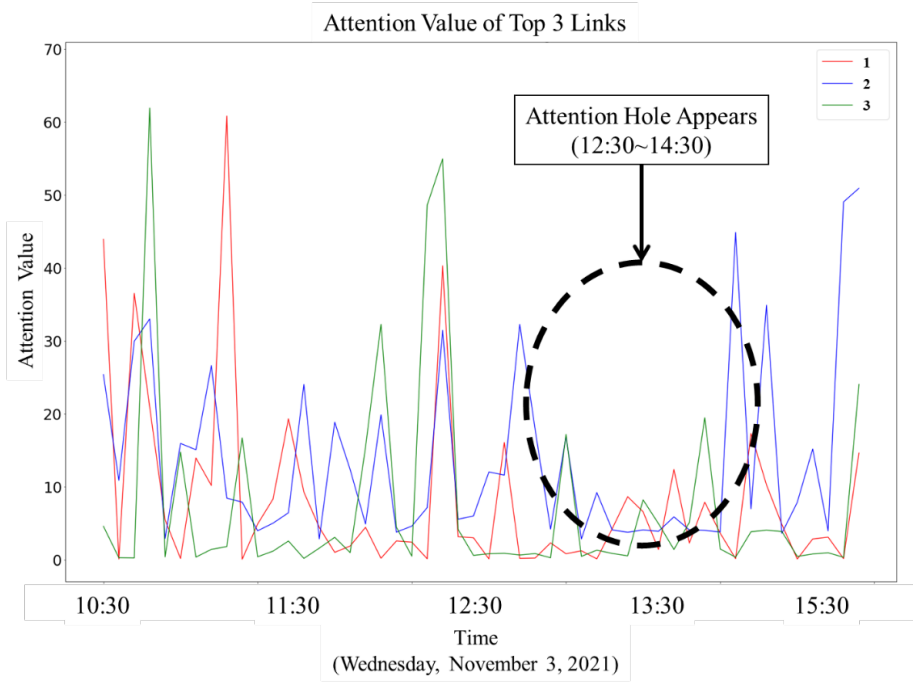


Figure 5.8: Attention value of top 3 links of Gangnam-gu on Wednesday, November 3, 2021, around the lunchtime

5.2 Considering Road Network Heterogeneity and Traffic Volume

We evaluated the performance of the model from two perspectives. The first is the predictive accuracy for the average link, and the second is the predictive accuracy for the link with a high index value. For the second part of the standard, we verified the performance of each model using two widely used concepts in prediction models, precision and recall.

Precision and recall are two important performance metrics often used in prediction models. Precision refers to the fraction of true positive predictions out of all positive predictions made by the model, essentially measuring the model's accuracy in identifying positive cases. On the other hand, Recall measures the fraction of true positive predictions out of all actual positive cases, assessing the model's ability to identify all relevant cases. Balancing these two metrics is crucial for a robust prediction model, as a high precision with low recall indicates that the model is overly conservative and misses many relevant cases. In contrast, a high recall with low precision suggests that the model produces many false positives.

Although the concepts of precision and recall do not exactly match the situation here, they can be applied similarly. Our main goal is to find links with high IANI values. Therefore, we can evaluate the model based on how well it predicts the top-k values. Accurately predicting the actual top-k is similar to recall while predicting the model's top-k well is similar to precision. In the example shown in Figure 5.9, Prediction 1 has a perfect match for the top-3 predicted values with the actual ones, so the precision is 100%. However, the recall is 0%. On the other hand, Prediction 2 has perfectly matched the top-3, so the ranking accuracy from this perspective is 100%, and the recall is 50%.

We will use such a measure to compare the models.

True	**A**	**B**	**C**	D	E	F	G
	100	**80**	**60**	50	40	30	25
Prediction 1	E	F	G	D	**A**	**B**	**C**
	40	30	25	20	**20**	**20**	**15**
Prediction 2	**A**	**B**	**C**	D	E	F	G
	50	**40**	**30**	5	5	0	0

Figure 5.9: Example to introduce precision and recall (Let the true is trying to judge the top-3 samples)

In such problems, recall is more important. It is more crucial to accurately predict the true high values rather than just predicting what the model thinks are high values. Although the model without attention loss has better performance up to the top 20, the performance beyond that is better with attention loss. Therefore, different strategies can be used depending on the situation. If there is a strategy applicable to more than 30 links, it is appropriate to use a model with attention loss. On the other hand, for strategies involving fewer links, using a model without attention loss may be more suitable.

Incorporating heterogeneous road hierarchies has proven to be effective, as shown in Table 5.3. Based on the MAPE criterion, it demonstrated a 10.3% reduction in prediction error. To determine the physical meaning of this indicator, it should be calculated using the MAPE, which represents the accuracy of predicting the total future congestion. Through this, the effectiveness of road heterogeneity has been validated. Here, the research successfully improved the ASTGCN.

We could incorporate the limited traffic volume into our model based on the structure in Figure 4.13. The MAPE and its standard deviation, when including traffic volume, are presented in Table 5.4. The MAPE and standard deviation

Table 5.3: MAPE and its standard deviation for homogeneous and heterogeneous network

	Homogeneous Network (ASTGCN)	Heterogeneous Network (AST-HGCN)
MAPE	13.83%	**12.40%**
Std. Dev.	0.20%	0.28%

decreased, indicating an overall model performance improvement.

Table 5.4: MAPE and its standard deviation for without and with volume data

	Without Volume Data (AST-HGCN)	With Volume Data (AST-HGCN)
MAPE	12.40%	**12.14%**
Std. Dev.	0.28%	**0.21%**

Surprisingly, the same phenomenon occurred in precision and recall errors. Both errors showed a decrease when the volume data was included. Including volume data as a new dimension appears to enhance our understanding of traffic patterns.

Table 5.5: Precision and recall error in the case of the model with and without traffic volume decoder

	Precision Error		Recall Error	
Num of top-k	Without Volume Data	With Volume Data	Without Volume Data	With Volume Data
100	61.72%	**59.24%**	54.44%	**54.21%**
75	64.40%	**61.82%**	55.90%	**55.68%**
50	68.39%	**65.66%**	57.66%	**57.44%**
30	73.05%	**70.15%**	59.55%	**59.35%**
20	75.94%	**72.96%**	60.69%	**60.49%**
10	81.01%	**77.88%**	61.77%	**61.54%**
5	86.44%	**82.84%**	61.41%	**61.11%**
3	94.51%	**89.54%**	60.96%	**60.61%**
1	134.32%	**126.23%**	60.37%	**59.95%**

5.3 The Result of Implementing Attention Loss and its Guidelines

The overall performance of a model with attention loss has decreased compared with a model without attention loss. The MAPE of AST-HGCN increased from 12.40% to 12.50%, which is a negligible level when considering the benefits obtained from interpretability. We additionally observed the precision and recall for a link with a high-valued index.

Table 5.6: Precision and recall error in the case of the model with and without attention loss

	Precision Error		Recall Error	
Num of top-k	Without Att. Loss	With Att. Loss	Without Att. Loss	With Att. Loss
100	**55.68%**	61.72%	56.06%	**54.44%**
75	**57.93%**	64.40%	57.26%	**55.90%**
50	**60.94%**	68.39%	58.56%	**57.66%**
30	**64.33%**	73.05%	59.67%	**59.55%**
20	**66.88%**	75.94%	**60.16%**	60.69%
10	**73.63%**	81.01%	**60.45%**	61.77%
5	**79.12%**	86.44%	**59.86%**	61.41%
3	**81.46%**	94.51%	**59.38%**	60.96%
1	**87.19%**	134.32%	**59.18%**	60.37%

Based on the experimental results, we were able to confirm that the model including attention loss exhibited superior performance in predictions beyond the top 30. As the main focus of this problem is actually to accurately predict

high values, we can assert that the model incorporating attention loss offers greater practical effectiveness.

As described in Chapter 4.2.3, when the number of links increases, the model fails to properly assign stronger attention values to specific links, which the model treats as a critical link. To address this issue, we introduced the attention penalty. In this chapter, we examined whether the attention values actually concentrate on a random link as the number of links increases.

As the number of links in a model expands, it is more likely for the attention value to be concentrated. This problem is significantly more challenging for the case of Seoul than the case of Gangnam District. In Gangnam District, the model needs to find the link to focus the attention value among 228 links, whereas, for Seoul, it needs to find the link among 5,068 links. This can be considered about 20 times more difficult. Due to this, situations, where all attention values are concentrated on a single link, occur often in the training process with Seoul. This deviates significantly from the actual scenario. However, by applying attention loss, the model can overcome this issue.

Table 5.7 demonstrates the severity of the phenomenon of extreme concentration. The table presents the concentration loss values collected randomly sampling 50 to 3,000 links without the attention penalty during training. As the number of links increases, the concentration loss increases and the ratio compared to the case where all attention is focused on a single link also converges to around 43.46% with 3,000 links.

The concentration loss counts indicate how many models trained with a specific number of links exceeded a certain ratio. For example, 40% of "Over 12.5%" in the case of 1,000 links means that 40% of the models had a concentration ratio higher than 12.5%. This ratio also increases as the number of links increases.

Table 5.7: Concentration loss and concentration loss count by the number of links

Number of Links	Concentration Loss			Concentration Loss Count			
	Mean	Std. Dev.	Ratio	Over 6.25%	Over 12.5%	Over 25.0%	Over 50.0%
50	1.27	0.14	2.55%	0%	0%	0%	0%
64	1.82	1.93	2.85%	5%	5%	0%	0%
100	1.53	0.27	1.53%	0%	0%	0%	0%
128	6.35	5.09	4.96%	45%	0%	0%	0%
256	9.43	7.66	3.68%	25%	0%	0%	0%
500	33.23	16.74	6.65%	50%	5%	0%	0%
750	105.46	118.18	14.06%	80%	25%	15%	5%
1000	123.89	63.80	12.39%	90%	40%	5%	0%
1500	337.01	306.63	22.47%	100%	75%	20%	5%
2000	518.34	464.15	25.92%	90%	70%	35%	15%
2500	817.18	788.39	32.69%	100%	80%	30%	20%
3000	1303.84	1007.25	43.46%	100%	95%	55%	35%

Figure 5.10 illustrates the key findings from Table 5.7. However, it is not easy to establish a clear upper threshold for the ratio to determine what is considered normal. Determining whether assigning 6.65% of attention to a single link out of 500 links is appropriate or inappropriate is challenging. Nevertheless, as the number of links increases, it is natural for the ratio to decrease. However, in actual training, the opposite occurs. Therefore, additional penalties should be applied to ensure proper training.

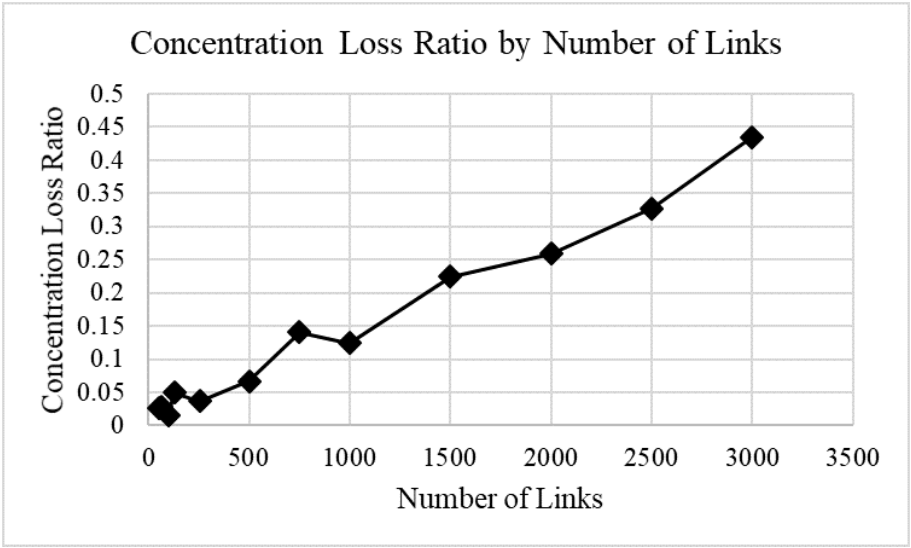


Figure 5.10: Concentration loss ratio by number of links

Chapter 6

Conclusion

Through this study, we addressed the problem of link impact index definition and its prediction, taking into account real-time, future prediction, and traffic network reflection. The inference time of the model developed in this study is within a few seconds using the NVIDIA A100 graphic card, guaranteeing real-time performance. By employing the developed model, the priority links of the Seoul road network can be identified. The results were analyzed based on various factors, such as the model's structure and the application of attention loss. Based on the analysis results, it is anticipated that we will be able to respond quickly to situations when a disaster or accident occurs in the future. Furthermore, by presenting a path detour strategy, it is possible to suggest a strategy that can drive close to the social optimum out of the current user equilibrium. The contributions of this paper can be summarized as follows:

This paper's contributions can be outlined in three aspects. First, we proposed a new index to identify the influence of links to networks in real-time. Second, we introduced a high-performance speed prediction model based on a

graph attention model by constructing a traffic flow-reflecting adjacency matrix value. Despite being a research product, the model's accuracy demonstrates its potential for various applications. Third, the research suggested various models upon considering the heterogeneity of the network or considering the attention loss. Based on this study, the location and implementation of countermeasures for speed reduction can be determined.

While conducting attention analysis and developing the model, it was recognized that the amount of change is more important than the attention value itself. We anticipate that focusing on the difference in attention values rather than the attention values themselves can yield more stable results. To this end, we aim to address various events by determining the average of the attention values derived from the previously trained model and indexing the amount of change here.

This study is significant as it simultaneously addresses priority node identification and speed reduction prediction. However, there are still areas that need further research. First, by directly obtaining accident and disaster data, it is necessary to identify the priority links during actual events and how these priority links change over time. We need to evaluate the strategies proposed by the model through case studies on more diverse rare events. Second, The most significant limitation is that the data used in this research is aggregated at the link level rather than the lane level. Therefore, we can only obtain aggregated data for vehicles that traveled on specific links for a 5-minute interval rather than individual vehicle movement data. The data is in the form of aggregated data rather than movement-level data. We hope to address this limitation in future research when more detailed data becomes available.

Bibliography

Afrin, T. and Yodo, N. (2020). A survey of road traffic congestion measures towards a sustainable and resilient transportation system. *Sustainability*, **12**(11), 4660.

Asif, M. T., Dauwels, J., Goh, C. Y., Oran, A., Fathi, E., Xu, M., Dhanya, M. M., Mitrovic, N., and Jaillet, P. (2013). Spatiotemporal patterns in large-scale traffic speed prediction. *IEEE Transactions on Intelligent Transportation Systems*, **15**(2), 794–804.

Aydin, N. Y., Duzgun, H. S., Heinimann, H. R., Wenzel, F., and Gnyawali, K. R. (2018). Framework for improving the resilience and recovery of transportation networks under geohazard risks. *International journal of disaster risk reduction*, **31**, 832–843.

Bell, M. G., Kurauchi, F., Perera, S., and Wong, W. (2017). Investigating transport network vulnerability by capacity weighted spectral analysis. *Transportation Research Part B: Methodological*, **99**, 251–266.

Chen, C., Li, K., Teo, S. G., Zou, X., Wang, K., Wang, J., and Zeng, Z. (2019). Gated residual recurrent graph neural networks for traffic prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 485–492.

Chen, D., Lin, Y., Li, W., Li, P., Zhou, J., and Sun, X. (2020). Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3438–3445.

Dawson, D., Shaw, J., and Gehrels, W. R. (2016). Sea-level rise impacts on transport infrastructure: The notorious case of the coastal railway line at dawlish, england. *Journal of Transport Geography*, **51**, 97–109.

Dong, C., Shao, C., Richards, S. H., and Han, L. D. (2014). Flow rate and time mean speed predictions for the urban freeway network using state space models. *Transportation Research Part C: Emerging Technologies*, **43**, 20–32.

Galimberti, C. L., Furieri, L., Xu, L., and Ferrari-Trecate, G. (2023). Hamiltonian deep neural networks guaranteeing nonvanishing gradients by design. *IEEE Transactions on Automatic Control*, **68**(5), 3155–3162.

Gauthier, P., Furno, A., and El Faouzi, N.-E. (2018). Road network resilience: how to identify critical links subject to day-to-day disruptions. *Transportation research record*, **2672**(1), 54–65.

Ge, L., Li, H., Liu, J., and Zhou, A. (2019). Temporal graph convolutional networks for traffic speed prediction considering external factors. In *2019 20th IEEE International Conference on Mobile Data Management (MDM)*, pages 234–242. IEEE.

Gu, Y., Fu, X., Liu, Z., Xu, X., and Chen, A. (2020). Performance of transportation network under perturbations: Reliability, vulnerability, and resilience. *Transportation Research Part E: Logistics and Transportation Review*, **133**, 101809.

Guo, S., Lin, Y., Feng, N., Song, C., and Wan, H. (2019). Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 922–929.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jia, Y., Wu, J., and Du, Y. (2016). Traffic speed prediction using deep learning method. In *2016 IEEE 19th international conference on intelligent transportation systems (ITSC)*, pages 1217–1222. IEEE.

Jun, M.-J. (2020). The effects of polycentric evolution on commute times in a polycentric compact city: A case of the seoul metropolitan area. *Cities*, **98**, 102587.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, **18**(1), 39–43.

Kipf, T. N. and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kong, X., Xing, W., Wei, X., Bao, P., Zhang, J., and Lu, W. (2020). Stgat: Spatial-temporal graph attention networks for traffic flow forecasting. *IEEE Access*, **8**, 134363–134372.

Li, M. and Zhu, Z. (2021). Spatial-temporal fusion graph neural networks for traffic flow forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4189–4196.

Li, Y., Yu, R., Shahabi, C., and Liu, Y. (2017). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*.

Li, Z., Jin, C., Hu, P., and Wang, C. (2019). Resilience-based transportation network recovery strategy during emergency recovery phase under uncertainty. *Reliability Engineering & System Safety*, **188**, 503–514.

Liu, M., Gao, H., and Ji, S. (2020). Towards deeper graph neural networks. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 338–348.

Liu, Y., McNeil, S., Hackl, J., and Adey, B. T. (2022). Prioritizing transportation network recovery using a resilience measure. *Sustainable and Resilient Infrastructure*, **7**(1), 70–81.

Lu, Z., Lv, W., Cao, Y., Xie, Z., Peng, H., and Du, B. (2020). Lstm variants meet graph neural networks for road speed prediction. *Neurocomputing*, **400**, 34–45.

Ma, X., Tao, Z., Wang, Y., Yu, H., and Wang, Y. (2015). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies*, **54**, 187–197.

Ma, X., Dai, Z., He, Z., Ma, J., Wang, Y., and Wang, Y. (2017). Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction. *Sensors*, **17**(4), 818.

Mattsson, L.-G. and Jenelius, E. (2015). Vulnerability and resilience of transport systems–a discussion of recent research. *Transportation research part A: policy and practice*, **81**, 16–34.

Min, W. and Wynter, L. (2011). Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C: Emerging Technologies*, **19**(4), 606–616.

Nagy, A. M. and Simon, V. (2021). Improving traffic prediction using congestion propagation patterns in smart cities. *Advanced Engineering Informatics*, **50**, 101343.

Nguyen, H., Liu, W., and Chen, F. (2016). Discovering congestion propagation patterns in spatio-temporal traffic data. *IEEE Transactions on Big Data*, **3**(2), 169–180.

Pan, B., Demiryurek, U., and Shahabi, C. (2012). Utilizing real-world transportation data for accurate traffic prediction. In *2012 ieee 12th international conference on data mining*, pages 595–604. IEEE.

Pan, X., Dang, Y., Wang, H., Hong, D., Li, Y., and Deng, H. (2022). Resilience model and recovery strategy of transportation network based on travel od-grid analysis. *Reliability Engineering & System Safety*, **223**, 108483.

Park, C., Lee, C., Bahng, H., Tae, Y., Jin, S., Kim, K., Ko, S., and Choo, J. (2020). St-grat: A novel spatio-temporal graph attention networks for accurately forecasting dynamically changing road speed. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1215–1224.

Polson, N. G. and Sokolov, V. O. (2017). Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, **79**, 1–17.

Ramezani, M. and Geroliminis, N. (2015). Queue profile estimation in congested urban networks with probe data. *Computer-Aided Civil and Infrastructure Engineering*, **30**(6), 414–432.

Rao, A. M. and Rao, K. R. (2012). Measuring urban traffic congestion-a review. *International Journal for Traffic & Transport Engineering*, **2**(4).

Sun, Y., Jiang, G., Lam, S.-K., and He, P. (2021). Learning traffic network embeddings for predicting congestion propagation. *IEEE Transactions on Intelligent Transportation Systems*, **23**(8), 11591–11604.

Van Aken, S., Bešinović, N., and Goverde, R. M. (2017). Solving large-scale train timetable adjustment problems under infrastructure maintenance possessions. *Journal of rail transport planning & management*, **7**(3), 141–156.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., *et al.* (2017). Graph attention networks. *stat*, **1050**(20), 10–48550.

Wang, J. and Shi, Q. (2013). Short-term traffic speed forecasting hybrid model based on chaos–wavelet analysis-support vector machine theory. *Transportation Research Part C: Emerging Technologies*, **27**, 219–232.

Wang, J., Gu, Q., Wu, J., Liu, G., and Xiong, Z. (2016). Traffic speed prediction and congestion source exploration: A deep learning method. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 499–508. IEEE.

Wang, J., Chen, R., and He, Z. (2019). Traffic speed prediction for urban transportation network: A path based deep learning approach. *Transportation Research Part C: Emerging Technologies*, **100**, 372–385.

Wang, X., Ma, Y., Wang, Y., Jin, W., Wang, X., Tang, J., Jia, C., and Yu, J. (2020). Traffic flow prediction via spatial temporal graph neural network. In *Proceedings of the web conference 2020*, pages 1082–1092.

Wu, Y., Tan, H., Qin, L., Ran, B., and Jiang, Z. (2018). A hybrid deep learning based traffic flow prediction method and its understanding. *Transportation Research Part C: Emerging Technologies*, **90**, 166–180.

Yang, X., Liu, L., Li, Y., and He, R. (2016). Identifying critical links in urban traffic networks: a partial network scan algorithm. *Kybernetes*.

Yu, B., Yin, H., and Zhu, Z. (2017). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*.

Yu, B., Lee, Y., and Sohn, K. (2020). Forecasting road traffic speeds by considering area-wide spatio-temporal dependencies based on a graph convolutional neural network (gcn). *Transportation research part C: emerging technologies*, **114**, 189–204.

Yu, C., Yang, X., and Yun, M. (2014). Method of searching for critical links in traffic network based on link redundancy. In *Transportation Research Board 93rd Annual Meeting*, pages 12–16.

Zhang, C., James, J., and Liu, Y. (2019a). Spatial-temporal graph attention networks: A deep learning approach for traffic forecasting. *IEEE Access*, **7**, 166246–166256.

Zhang, X., Huang, C., Xu, Y., and Xia, L. (2020). Spatial-temporal convolutional graph attention networks for citywide traffic flow forecasting. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1853–1862.

Zhang, Z., Li, M., Lin, X., Wang, Y., and He, F. (2019b). Multistep speed prediction on traffic networks: A deep learning approach considering spatio-temporal dependencies. *Transportation research part C: emerging technologies*, **105**, 297–322.

Zheng, C., Fan, X., Wang, C., and Qi, J. (2020). Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 1234–1241.

Zhou, F., Yang, Q., Zhong, T., Chen, D., and Zhang, N. (2020). Variational graph neural networks for road traffic prediction in intelligent transportation systems. *IEEE Transactions on Industrial Informatics*, **17**(4), 2802–2812.

Zou, Y., Hua, X., Zhang, Y., and Wang, Y. (2015). Hybrid short-term freeway speed prediction methods based on periodic analysis. *Canadian Journal of Civil Engineering*, **42**(8), 570–582.

국문초록

교통 혼잡은 도로 이동의 비효율성을 초래하는 장애물로, 통행 시간과 연료 이용을 증가시켜 경제적 및 환경적 비용을 초래한다. 이러한 문제의식에서 출발하여 본 연구는 도로 네트워크 내의 링크가 인접 네트워크에 미치는 영향을 파악해 미래 교통 상황에 더 지대한 영향을 끼치는 링크를 선별하는 것을 목표로 한다. 대한민국의 수도인 서울의 도심부 도로 네트워크에 중점을 둔 이 연구에서는 새롭게 정의한 Impact on Adjacency Network Index (IANI)를 기반으로 네트워크 규모의 속도 감소를 예측할 수 있는 모델을 개발했다. 이 모델은 교통 흐름과 도로 네트워크의 특성을 고려하도록 설계되었다. 교통 흐름의 특성을 반영하기 위해서 새롭게 개발된 인접 행렬이 활용되었으며, 서로 다른 위계를 갖는 연속류 및 단속류 흐름을 고려하여 도로 네트워크 특성을 반영하였다. 또한, 어텐션 값에 대한 손실 함수를 도입하여 그래프 어텐션 모델의 현실성과 예측 결과의 신뢰성을 향상하였다.

교통 흐름이 고려된 인접 행렬은 그래프 어텐션 모델과 함께 활용되었을 때 전통적인 거리 기반의 인접 행렬에 비해 향상된 성능을 보였다. 단속류를 구분한 경우에도 마찬가지로 IANI 예측값의 정확도가 상승하는 것을 확인할 수 있었다. 현실성있는 어텐션 값을 위해 손실 함수에 어텐션 값을 추가한 경우 예측 성능 자체는 악화된다. 하지만 이에 비해 현실적인 주요 링크 선별에 중요한 recall의 값은 상승하였기에 이점이 더 많다고 할 수 있다. 본 연구에서 제시한 모델은 다양한 교통 시나리오에서 실시간 주요 링크 선별을 통한 대응 가능성을 보여준다. 이 모델의 결과는 신호 최적화 및 도로 확장과 같은 교통 전략 측면에서 사용될 것으로 기대된다.

주요어: 그래프 어텐션 모델, 속도 감소 예측, 네트워크 영향력, 이질적 도로 네트워크, 어텐션 손실 함수

학번: 2018-25029