BMC Public Health

RESEARCH Open Access

A novel framework for inferring dynamic infectious disease transmission with graph attention: a COVID-19 case study in Korea

Minji Lee[1], Heejin Choi[1] and Chang Hyeong Lee[1*]

Abstract

Introduction Epidemic modeling is crucial for understanding and predicting infectious disease spread. To capture the complexity of real-world transmission, dynamic interactions between individuals with spatial heterogeneity must be considered. This modeling requires high-dimensional epidemic parameters, which can lead to unidentifiability; therefore, integrating various data types for inference is essential to effectively address these challenges.

Methods We introduce a novel hybrid framework, Multi-Patch Model Update with Graph Attention Network (MPUGAT), that combines a multi-patch compartmental model with a spatio-temporal deep learning model. MPUGAT employs a GAT (Graph Attention Mechanism) to transform static traffic matrices into dynamic transmission matrices by analyzing patterns in diverse time series data from each city.

Results We demonstrate the effectiveness of MPUGAT through its application to COVID-19 data from South Korea. By accurately estimating time-varying transmission rates, MPUGAT outperforms traditional models and aligns with actual policies such as social distancing.

Conclusion MPUGAT offers a novel approach for effectively integrating easily accessible, low-dimensional, non-epidemic-related data into epidemic modeling frameworks. Our findings highlight the importance of incorporating dynamic data and utilizing graph attention mechanisms to enhance accuracy of infectious disease modeling and the analysis of policy interventions. This study underscores the potential of leveraging diverse data sources and advanced deep learning techniques to improve epidemic forecasting and inform public health strategies.

Keywords Deep learning, Multi-patch model, Epidemic modeling, Compartment model, Transmission matrix, Contact pattern

Introduction

Infectious diseases have been a recurring and persistent threat throughout human history, causing immense economic and human losses. COVID-19 pandemic, with its millions of cases and millions of deaths, underscores this stark reality. In an increasingly interconnected and complex world, the emergence and global spread of infectious diseases have been accelerating [1]. Modeling these outbreaks is crucial for proactive prevention and mitigation. Epidemic dynamic models enable the timely implementation of public health interventions and the efficient allocation of resources [2]. Since the seminal work by Kermack and McKendrick in 1927 [3], infectious disease modeling has significantly evolved to analyze a wide array of infectious diseases.

The transmission of many infectious diseases is driven by human contact, where the transmission rate is influenced by the frequency of contacts and the

*Correspondence:
Chang Hyeong Lee
chlee@unist.ac.kr
[1] Department of Mathematical Sciences, Ulsan National Institute of Science and Technology, UNIST-gil 50, Ulsan 44919, Republic of Korea

cumulative transmission risk over time. Thus, the way contact patterns are incorporated into epidemiological models greatly influences their predictive accuracy [4–6]. Early models assumed a homogeneous population with uniform transmission within well-mixed compartments [7]. However, real-world transmission patterns often depend on social structure [8, 9], age [10], spatial distribution [11], time [12], and among other factors. Specifically, the spatial dependency of disease transmission is evident through spatial autocorrelation, where nearby regions exhibit similar patterns of disease incidence. For example, analyses of COVID-19 transmission in South Korea demonstrated significant spatial clustering around outbreak locations [13, 14]. This spatial pattern underscores the importance of incorporating geographic heterogeneity into epidemiological models. Consequently, incorporating heterogeneity into models is essential for improving their predictive and explanatory power. Several mathematical models have been utilized to capture spatial heterogeneity in disease spread. There are three main approaches-mathematical models, data-driven models, and hybrid models that integrate both.

Mathematical models, particularly multi patch models (meta population models), provide a powerful framework to capture heterogeneity in disease transmission [15, 16]. Dividing heterogeneous individuals into well-mixed patches-based on regions [17, 18] or other criteria [19, 20]-allows these models to capture diverse contact patterns and transmission rates between patches through a transmission matrix. Additionally, Agent-Based Models (ABMs) [21] and network modeling [18] represent cities as nodes connected by geographical links, accounting for structure variations. Previous studies have derived epidemic parameters, such as transmission matrices for multi-patch models and network structures for ABMs and network modeling, using data from traffic, air travel, distance, and population [21–24]. Additionally, statistical techniques, such as inverse methods and least-squares approaches, have been employed for parameter estimation [25]. While mathematical models provide valuable insights into disease dynamics and inform policy decisions, constructing epidemic parameters remains a complex task. Obtaining spatial-temporal high-dimensional data is difficult [26, 27], and the quality of this data significantly impacts the accuracy of the epidemic parameters [28, 29]. When statistical methods are employed to infer epidemic parameters, the high dimensionality can lead to unidentifiability, where different parameter values yield similar observational dynamics [30, 31]. This underscores the need for integrating diverse data sources for reliable inference.

Data-driven models offer flexible frameworks that can better fit data and integrate diverse sources without the need for obtaining epidemic parameters. With the increasing availability of data, deep learning models have gained significant attention among data-driven approaches [32]. In spatial epidemic modeling, many methods utilize Graph Neural Networks (GNNs), where cities are represented as nodes and inter-city connections as edges. These models are often combined with temporal deep learning techniques, such as Long Short-Term Memory (LSTM) networks, to capture sequential dependencies [33–35]. However, their black-box nature makes interpretation difficult and can sometimes produce results that contradict established epidemiological principles. To overcome these limitations, recent studies have focused on hybrid models that integrate mathematical models with deep learning approaches. One of the most prominent hybrid models, Physics-Informed Neural Networks (PINNs), incorporates physical constraints into the loss function of neural networks to solve ordinary differential equations (ODEs). While PINNs have been extensively studied in single-patch settings [36–38], their computational cost increases exponentially with model complexity, limiting their applicability to multi-patch scenarios. Given that epidemic data is typically collected on a daily basis, many studies have utilized deep learning models to predict daily epidemic parameters and integrate them into traditional mathematical models. Due to their computational efficiency and ease of implementation, this approach has been widely applied in both single-patch [39–41] and multi-patch [42–45] models.

Mathematical, data-driven, and hybrid models have significantly contributed to epidemic modeling. However, there are still challenges in estimating the high-dimensional epidemic parameters required to build interpretable spatial epidemic models. To address these challenges, we propose the Multi-Patch Model Update with Graph Attention Network (MPUGAT), a hybrid framework for inferring dynamic infectious disease transmission from diverse data sources. This model leverages the attention mechanism of Graph Attention Networks (GATs), which dynamically learn node connections through attention scores (weighted edges) rather than relying on predefined links. Previous studies have utilized GATs to enhance spatiotemporal pattern recognition and long-term forecasting [35, 45]. However, these approaches primarily focus on prediction rather than inferring high-dimensional spatiotemporal epidemic parameters. Notably, [43] introduced an interesting method that updates the transmission matrix using GNNs based on multiple time series data. However, static traffic data can only infer a static transmission matrix, and the deep learning model size increases quadratically with the number of cities.

Our key innovation lies in MPUGAT's ability to generate a dynamic transmission matrix by utilizing the graph attention mechanism with traffic and various time series data. It also maintains model size efficiency even as the number of cities increases. MPUGAT achieves this by building a multi-patch compartment model that distinguishes between populations based on residency status and current location [17], along with guiding attention scores in a desired direction [46]. To the best of our knowledge, this is the first application of a graph attention mechanism for inferring and extending epidemic parameters. Figure 1 provides an overview of MPUGAT. The model takes time series data from each city, along with static either dynamic traffic data, as input. The deep learning component, which incorporates Long Short-Term Memory (LSTM) and the Graph Attention Mechanism (GAT), processes this data to estimate a dynamic transmission matrix, which is subsequently used to solve the ODE of the multi-patch SEIQ model. This framework provides valuable insights into epidemic dynamics and enables the analysis of various policy interventions.

The remainder of this paper is organized as follows. The Methods section details the methodology, including the multi-patch SEIQ model, the spatio-temporal deep learning model, and then proposes our hybrid framework integrating these two models. The Case Study section presents a case study using COVID-19 data from South Korea to demonstrate the efficacy of MPUGAT. Finally, the Conclusion and Discussion section discusses the findings, limitations, and future research directions.

Methods

Multi-patch SEIQ compartment model

We employ a mathematical compartment model to integrate heterogeneous transmission patterns into the epidemic modeling. Specifically, we propose a multi-patch SEIQ model by incorporating an exposed compartment and adapting the patch structure to represent geographical regions (Fig. 2A).

In the model, the compartments include Susceptible S (individuals who can be infected), Exposed E (individuals in the latent period of infection), Infected I (individuals who are currently infected and can infect others), and Quarantine and Recovered Q (individuals who are quarantined and then recovered). The total population size, denoted by N, remains constant and is given by $N:=S+E+I+Q$. Furthermore, the population is geographically distributed across multiple cities, as indicated by subscripts, based on residential registration. This classification is used because most data is reported according to residents' registered addresses. Thus, we need to infer the exact contact patterns between residents of different cities. As shown in Fig. 2B, residents of city l and city k can interact in city j, and similarly, residents from any city can interact without restrictions on location. This interaction is captured by the contact matrix $C_{lk}(t)$, which is used in the model as a component of the transmission matrix. Through this contact matrix, the dynamics of infectious diseases among the cities are interconnected and influenced by one another. The dynamics of this multi-patch SEIQ model, incorporating

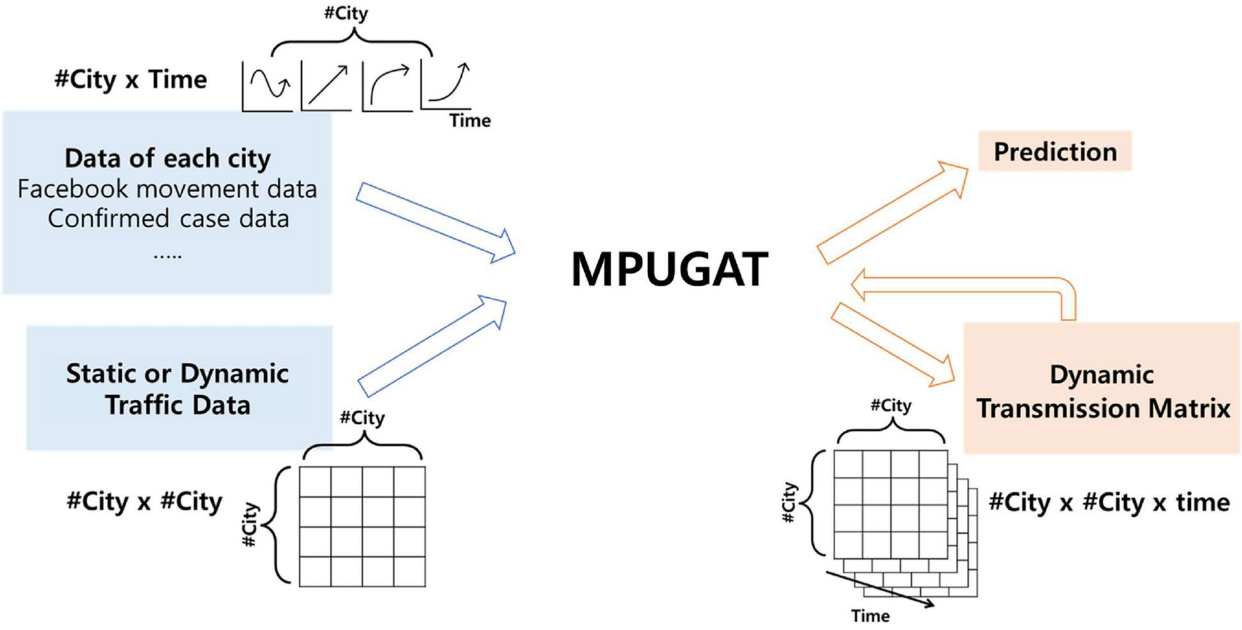


Fig. 1 Overview of the proposed MPUGAT. The time series data from each city is processed through a deep learning model, updating traffic data into a dynamic transmission matrix within a multi-patch compartment model

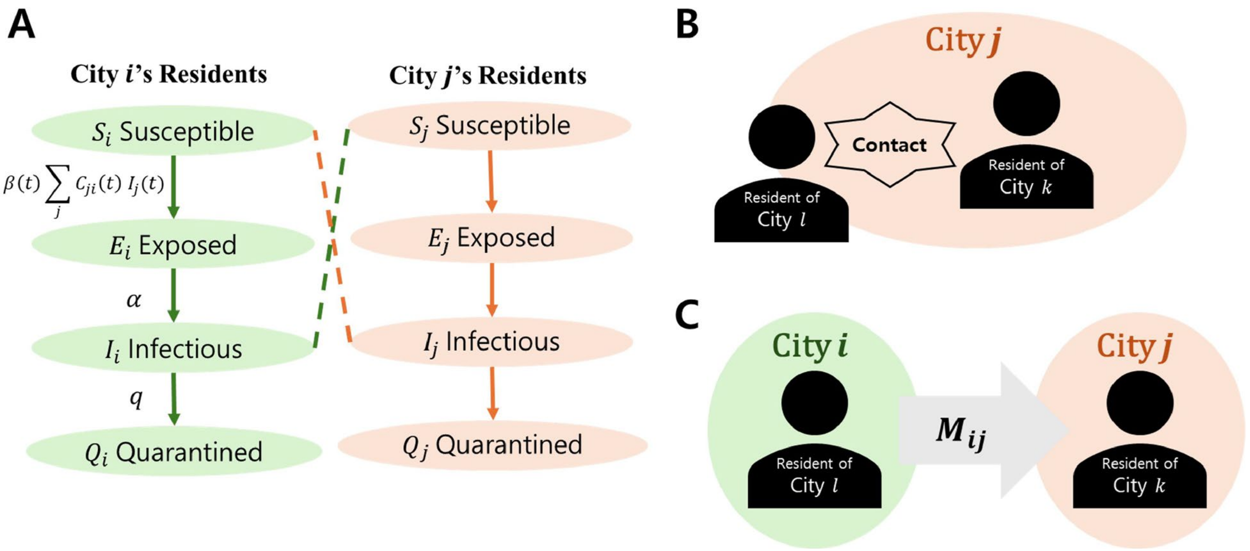


Fig. 2 **A** Diagram of the multi-patch SEIQ compartment model. Solid lines indicate state change within a city, while dashed lines represent influences between cities. **B** Residents can contact each other regardless of city boundaries. **C** The movement data M_{ij} reflects the strength of interaction (movement) between residents traveling from City i to City j, encompassing various residents

the contact matrix between residents, can be described by the following Eq. (1) and the description of the parameters is provided in Table 1.

$$\begin{cases}\dot{S}_i(t) = -\beta(t)\left(\sum_j \frac{C_{ji}(t)I_j(t)}{N_j}\right)\frac{S_i(t)}{N_i(t)} \\ \dot{E}_i(t) = \beta(t)\left(\sum_j \frac{C_{ji}(t)I_j(t)}{N_j}\right)\frac{S_i(t)}{N_i(t)} - \alpha E_i(t) \\ \dot{I}_i(t) = \alpha E_i(t) - qI_i(t) \\ \dot{Q}_i(t) = qI_i(t)\end{cases} \tag{1}$$

We assume a homogeneous transmission rate per contact $\beta(t)$ across all cities, reflecting the absence of discernible differences based on registered residence. The transmission matrix T is defined as the product of the time-varying transmission rate $\beta(t)$ and the contact matrix $C_{ij}(t)$ as follows:

$$T_{ij}(t) = \beta(t) \times C_{ij}(t)$$

Here, each element of the matrix $T(t)$ denotes the transmission rate from residents in city i to residents in city j at time t.

Decomposition of contact patterns

The ability of individuals to interact across cities without locational constraints introduces complexity in inferring

Table 1 Compartment and parameter descriptions

	Description
Compartment	
$S_i(t)$	Number of susceptible residents in city i at time t
$E_i(t)$	Number of exposed residents (latent period) in city i at time t
$I_i(t)$	Number of infected residents in city i at time t
$Q_i(t)$	Number of quarantined and then recovered residents in city i at time t
N_i	Total population of city i
Parameter	
$\beta(t)$ [1]	Transmission rate per contact at time t
$C_{ij}(t)$ [2]	Adequate contacts between residents of city i and residents of city j at time t
$1/\alpha$	Latent period
$1/q$	Mean duration of case confirmation

[1] The transmission rate $\beta(t)$ is time-dependent to reflect changes in public health measures and social behavior

[2] $C_{ij}(t)$ represents the contact matrix that captures spatial interaction between cities

the contact matrix C_{ij}. While this matrix is often inferred from a single data source, such as movement data (as discussed in the introduction), it is crucial to recognize, as illustrated in Fig. 2C, that the volume of movement M_{ij} from city i to city j differs from the contact C_{ij}; the movement M_{ij} also encompasses interactions between residents from various cities, not just city i and city j.

To improve the accuracy of the inference of the dynamic transmission matrix from static movement data, we propose a decomposition of the contact matrix $C(t)$ into a resident ratio matrix $A(t)$ and a movement matrix M. The matrix $A_{ij}(t)$ represents the ratio of residents from city i among the population in city j at time t, capturing the residential distribution dynamics. On the other hand, M_{ij} denotes the movement data that quantifies the volume of movement from city i to city j.

The contact between residents of city i and city j can be described as the sum of decomposed contacts across all cities k. According to the mass action principle, this is expressed as the product of the number of residents from i and j present in each city k. To calculate the number of residents from city i present in city k at time t, we decompose the contact matrix using the movement data. The number of residents from i in city k is represented as the sum of residents from i moving into city k from various cities, including city i itself. Our assumption is that the mobility between cities is proportionally distributed based on the residency composition of each city. For example, if 30% of the current population in a city consists of residents originally from another city, we assume that 30% of movements originating from that city are attributed to residents from the other city. To formalize this, we introduce $A_{ij}(t)$, which denotes the proportion of residents from city i within the population of city j at time t. This matrix satisfies the condition $\sum_i A_{ij} = 1$. Using this, the expression for C_{ij} can ultimately be represented as follows:

$$C_{ij}(t) = \sum_k \Big(\big(\text{Number of residents from city } i \text{ in city } k\big) \times \big(\text{Number of residents from city } j \text{ in city } k\big)\Big) \tag{2}$$

$$= \sum_k \left(\sum_l \left(A_{il}(t) M_{lk}\right) \times \sum_l \left(A_{jl}(t) M_{lk}\right) \right) \tag{3}$$

In this context, the movement matrix M can be either static or dynamic. Considering cases where obtaining high-dimensional data may be difficult, we assumed that the resident ratio matrix $A(t)$ is time-dependent and the movement matrix M is constant. If a dynamic movement matrix $M(t)$ is used, we expect more accurate inferences to be possible. We infer the resident matrix $A(t)$ using various types of time series data from each city. A higher $A_{ij}(t)$ indicates significant population mixing between city i and city j. If cities are homogeneous, their similarity can be shown in non-epidemic data in each city, such as mobility patterns, social media, and card usage information [47–49]. However, this relationship is nonlinear and involves highly complex, hidden patterns that may manifest as both causes and outcomes across various datasets. To estimate these hidden similarities, we employ the graph attention mechanism introduced in the Introduction. The decomposition of the contact matrix into a resident ratio matrix and movement matrix offers two key advantages. First, by leveraging static movement data, we can simultaneously infer the dynamic $A_{ij}(t)$ using deep learning across diverse datasets. Second, rather than inferring $C_{ij}(t)$, which lacks an upper bound, we can simultaneously and efficiently estimate $A(t)$ and $\beta(t)$, both of which are constrained within the unit interval, $0 \leq A(t), \beta(t) \leq 1$.

Spatio-temporal deep learning model

We infer the epidemic parameters for a future period h using a spatio-temporal deep learning model. This approach is adopted because, during the MPUGAT training process, we leverage a multi-patch compartment model to predict the future state value, $state(t) = \{S_i(t), E_i(t), I_i(t), Q_i(t)\}$. We estimate the $\beta(t) \in \mathbb{R}$ and $A(t) \in \mathbb{R}^{(C\times C)}$ using time-series data $x(t) \in \mathbb{R}^{(F\times C)}$, where C represents the number of cities and F corresponds to the number of input features. Importantly, this time-series data is not restricted to epidemic-related variables but can include any datasets that capture the distinctive characteristics of each city. The input data is represented as a graph $\mathbb{G}(t) = (\mathscr{N}(t), \mathscr{E}(t))$, where the nodes $\mathscr{N}(t) \in \mathbb{R}^{(F\times C\times w)}$ consist of time-series data over an input window w, constructed using a sliding window algorithm. Each node corresponding to city i, denoted as $\mathscr{N}_i(t) = \{x_i(t), x_i(t-1), \ldots, x_i(t-w+1)\} \in \mathbb{R}^{(F\times w)}$, contains the time-series data specific to city i. The edges $\mathscr{E}(t) \in \mathbb{R}^{(C\times C)}$ capture the connectivity between cities, with the initial edge weights set to one.

While the use of deep learning to predict epidemic parameters has been widely explored [39–45], we extend their application further. Specifically, we aim to infer resident ratio matrix, $A(t) \in \mathbb{R}^{(C\times C)}$ from the time series data of each city, $x(t) \in \mathbb{R}^{(F\times C)}$. Directly predicting A from data in the form of $(C \times C \times T)$ presents two main challenges. First, acquiring such high-dimensional data is often impractical. Second, inferring higher-dimensional structures from lower-dimensional datasets offers greater flexibility and practicality for real-world applications. To address this dimensionality problem, we adopt the attention mechanism inherent in the Graph Attention Network (GAT) architecture (see Fig. 3).

Initially, node $\mathscr{N}(t)$ is employed to predict future node dynamics using an LSTM over a fixed period h. We employ LSTM for temporal modeling, given its proven effectiveness in handling sequential data, such as city-level time series. In the context of epidemics, the initial information is crucial, and the LSTM memory cell enables the model to propagate information related to the early stages of the epidemic over long distances. The LSTM replaces the GAT in node embedding. Next, we learn a function to compute the attention score values, which capture the specific similarity between the LSTM embedding vectors of two nodes. At this stage, the attention score function is trained such that the attention scores represent the resident ratio matrix $A(t)$, which quantifies the proportion of residents from city i within the population of city j. The attention mechanism enables dynamic learning of relationships as the model adapts to new information, thereby enhancing its ability to reflect changes in the epidemic's dynamics over time. These coefficient values assign weights to each node's information, allowing for the aggregation of node data. Through this aggregated information, we estimate additional epidemic parameters, β. Importantly, the LSTM focuses on the temporal patterns within each city's data, while the GAT handles the spatial dependencies between cities. By combining LSTM and GAT, our model effectively captures both temporal and spatial dependencies, leveraging the strengths of each architecture. Our model can be summarized as shown in Fig. 4.

The detailed algorithm of the deep learning model can be found in Algorithm 1. To evaluate our approach, we developed several models, which can be reviewed in Table 4. The Deep Learning Model uses only LSTM and MLP to derive the *state*. In contrast, the MPUGAT, MPGAT, and Simple Hybrid Model utilize both LSTM and GAT to extract the epidemic parameters $\beta(t)$ and $A(t)$ of the multi-patch compartment model. These epidemic parameters are then employed in the compartment model to derive the *state* values. The next section will address this aspect in detail.

Algorithm 1 Detailed Algorithm for Deep Learning Model with LSTM and GAT

```
Parameters: $F$ (features), $w$ (window size), $h_{\mathrm{LSTM}}$ (hidden units), $C$ (number of cities), $h_{\mathrm{GAT}}$ (n heads of GAT), $l$ (num of LSTM layers), $a, b$ (learnable parameter), $I$ (identity matrix)

Input Graph $G = (\mathscr{N}, \mathscr{E})$, where $\mathscr{N}(t) \in \mathbb{R}^{F \times C \times w}$, $\mathscr{E} \in \mathbb{R}^{C \times C}$
Output $\beta(t) \in \mathbb{R}$, $A(t) \in \mathbb{R}^{C \times C}$

Deep Learning Model (LSTM)
for $i = 1$ to $l$ do
    $X_i(t) \leftarrow \mathrm{LSTM}(\mathscr{N}_i(t)) \in \mathbb{R}^{h_{\mathrm{LSTM}} \times w}$        ▷ LSTM processes temporal data
    $X_i(t) \leftarrow \mathrm{Tanh}(X_i(t))$        ▷ Apply activation function
end for
$X_i(t) \leftarrow X_i(t)[:, -1] \in \mathbb{R}^{h_{\mathrm{LSTM}}}$        ▷ Use last time step
$X(t) = (X_1(t), X_2(t), \ldots, X_C(t)) \in \mathbb{R}^{C \times h_{\mathrm{LSTM}}}$

MPUGAT, MPGAT, Simple Hybrid Model (LSTM+GAT)
$X_1(t) \leftarrow \mathrm{MLP}_1(X(t)) \in \mathbb{R}^{C \times h_{\mathrm{GAT}} \times h_{\mathrm{LSTM}}}$
$X_2(t) \leftarrow \mathrm{MLP}_2(X(t)) \in \mathbb{R}^{C \times h_{\mathrm{GAT}} \times h_{\mathrm{LSTM}}}$        ▷ Node embedding
Normalize $X_1(t), X_2(t)$ over $h_{\mathrm{LSTM}}$
$E(t) \leftarrow X_1(t) \cdot X_2(t)^\top + a * I \in \mathbb{R}^{C \times C \times h_{\mathrm{GAT}}}$
$X_3(t) \leftarrow \mathrm{Mean}(E \cdot X_1(t), axis = -1, -2, -3) \in \mathbb{R}$        ▷ Message aggregation
$\beta(t) \leftarrow \mathrm{sigmoid}(X_3(t))$
$A(t) \leftarrow \mathrm{softmax}(Mean(E, axis = -1)) \in \mathbb{R}^{C \times C}$        ▷ Attention mechanism
$C_{ij}(t) \leftarrow \sum_k (\sum_l (A_{il}(t) M_{lk}/b) \times \sum_l (A_{jl}(t) M_{lk}/b))$        ▷ $M$ is scaled by $b$

Total Learnable Parameters:
LSTM: $4 \times (h_{\mathrm{LSTM}} \times (F + h_{\mathrm{LSTM}}) + h_{\mathrm{LSTM}}) \times l$
GAT: $2 \times (h_{\mathrm{LSTM}} \times (h_{\mathrm{LSTM}} + h_{\mathrm{GAT}})) + 2$
```

Training algorithm

Leveraging the availability of daily confirmed case data for each city, we reformulate our multi-patch SEIQ model into a time-dependent mathematical model by applying finite difference methods. The resulting discretized system of equations, derived from Eq. (1), is presented as follows:

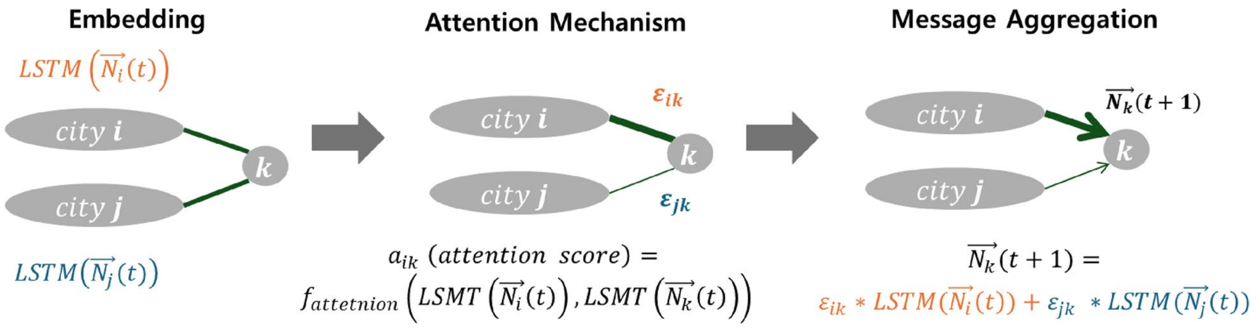

Embedding
Attention Mechanism
Message Aggregation
$LSTM(\overrightarrow{N_i}(t))$
city i
city j
k
$LSTM(\overrightarrow{N_j}(t))$
ε_{ik}
ε_{jk}
$\overrightarrow{N_k}(t+1)$
a_{ik} (attention score) =
$f_{attetnion}\left(LSMT\left(\overrightarrow{N_i}(t)\right), LSMT\left(\overrightarrow{N_k}(t)\right)\right)$
$\overrightarrow{N_k}(t+1) =$
$\varepsilon_{ik} * LSTM(\overrightarrow{N_i}(t)) + \varepsilon_{jk} * LSTM(\overrightarrow{N_j}(t))$

Fig. 3 A step-by-step visualization of how graph neural networks process city data. We learn the attention score matrix to become the resident ratio matrix A such that $A_{ij} = a_{ij}$

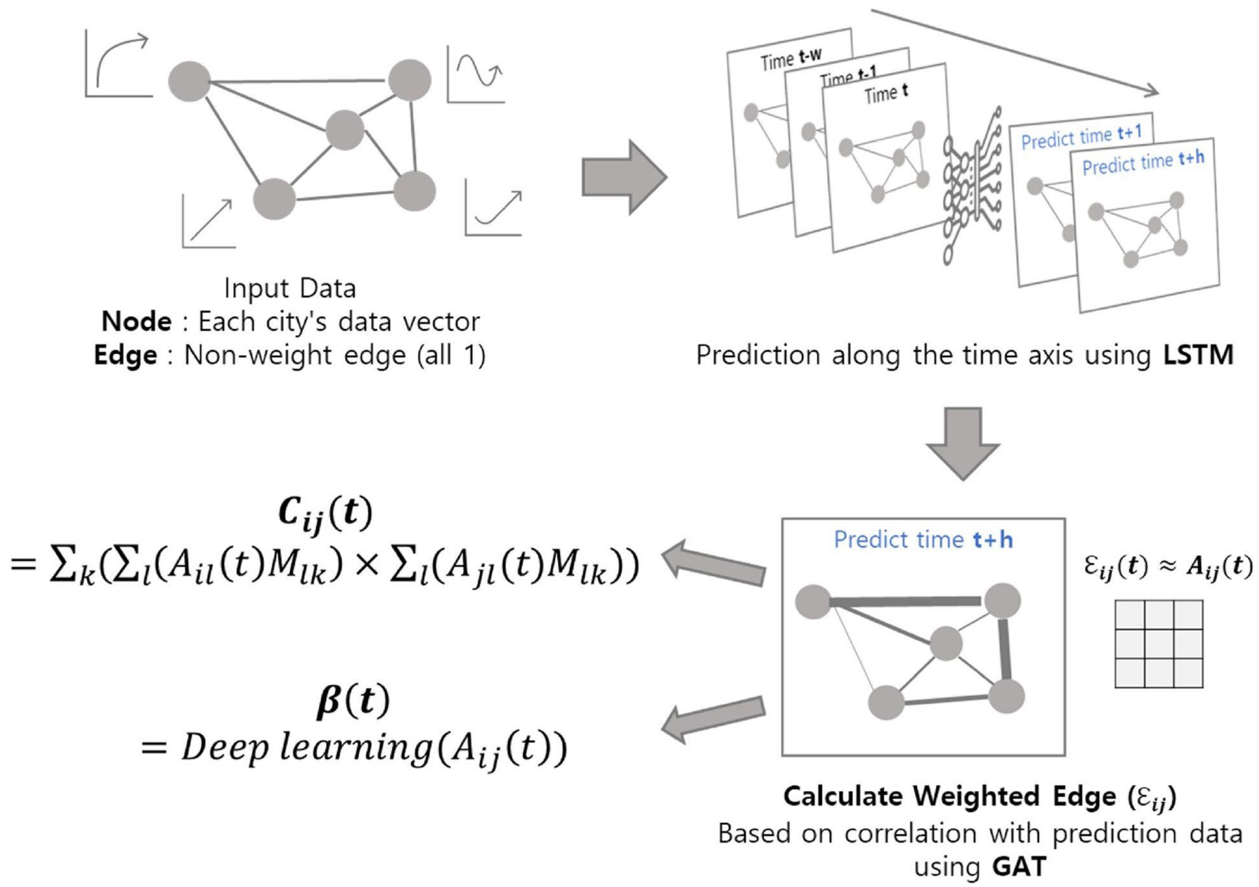

Input Data
Node : Each city's data vector
Edge : Non-weight edge (all 1)
Time t-w
Time t-1
Time t
Predict time t+1
Predict time t+h
Prediction along the time axis using **LSTM**
Predict time t+h
$\mathcal{E}_{ij}(t) \approx A_{ij}(t)$
Calculate Weighted Edge ($\mathcal{E}_{ij}$)
Based on correlation with prediction data using **GAT**
$C_{ij}(t) = \sum_k(\sum_l(A_{il}(t)M_{lk}) \times \sum_l(A_{jl}(t)M_{lk}))$
$\beta(t) = Deep\ learning(A_{ij}(t))$

Fig. 4 This figure illustrates the detailed structure of the deep learning algorithm in MPUGAT. The model infers a dynamic contact matrix and transmission rate, which are used to update static traffic data with time series data from each node (city)

$$\begin{cases} S_i(t+\Delta t) = S_i(t) - \Delta t \cdot \beta(t)\left(\sum_j \frac{C_{ji}(t)I_j(t)}{N_j}\right)\frac{S_i(t)}{N_i(t)} \\ E_i(t+\Delta t) = E_i(t) + \Delta t \cdot \left[\beta(t)\left(\sum_j \frac{C_{ji}(t)I_j(t)}{N_j}\right)\frac{S_i(t)}{N_i(t)} - \alpha E_i(t)\right] \\ I_i(t+\Delta t) = I_i(t) + \Delta t \cdot [\alpha E_i(t) - qI_i(t)] \\ Q_i(t+\Delta t) = Q_i(t) + \Delta t \cdot qI_i(t) \end{cases} \tag{4}$$

In this discretization scheme, Δt denotes the temporal increment between successive iterations, which is set to unity when utilizing daily reported case data. This iterative approach facilitates the prediction of future compartmental values based on the current system state and prevailing epidemiological parameters. Furthermore, by solving Eq. (4) in conjunction with the constraint $N_i = S_i(t) + I_i(t) + Q_i(t) + E_i(t)$, we can deduce the actual values of the epidemic states from the confirmed cases:

$$\begin{cases} S_i(t) = N - Q_i(t) - I_i(t) - E_i(t) \\ E_i(t) = \frac{I_i(t)(q+1) - I_i(t-1)}{\alpha} \\ I_i(t) = \frac{Q_i(t) - Q_i(t-1)}{q} \end{cases} \tag{5}$$

where $Q_i(t)$ is obtained from cumulative confirmed cases data provided by dataset [50]. The predicted future compartment value, denoted as $\hat{state}(t+h)$, is computed using the true current compartment value $state(t)$. First, Eq. (5) is used to calculate $state(t)$, and then, using the prediction parameters $\beta_i(t)$ and $C_{ij}(t)$ obtained from the deep learning model, $\hat{state}(t+1)$ is computed. The prediction is then recursively performed until $\hat{state}(t+h)$ is obtained, following Eq. (4). This process is illustrated in Fig. 5. We have set $h = 3$ to train the model based on the compartment state values two days ahead. This choice is made based on the fact that when $h = 1, 2$, using Eq. (4) to compute $\hat{state}(t+h)$ from the true values yields

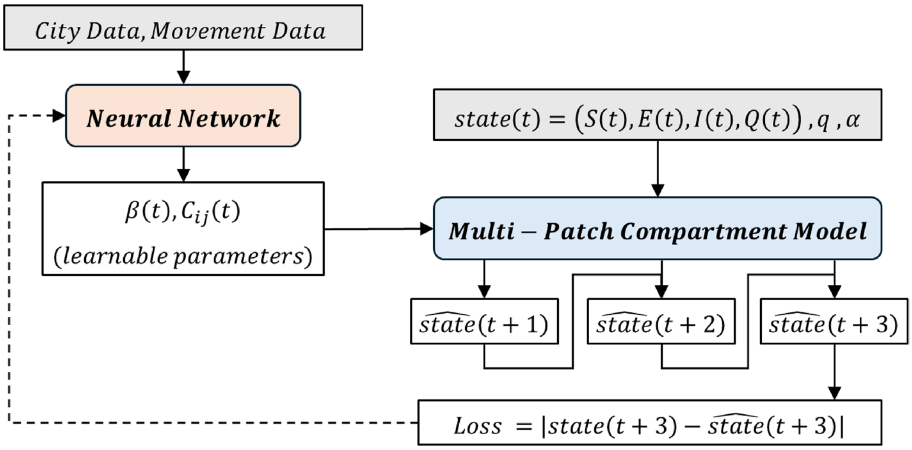

$City\ Data, Movement\ Data$
$\textbf{Neural Network}$
$\beta(t), C_{ij}(t)$
$(learnable\ parameters)$
$state(t) = \big(S(t), E(t), I(t), Q(t)\big), q, \alpha$
$\textbf{Multi − Patch Compartment Model}$
$\widehat{state}(t+1)$
$\widehat{state}(t+2)$
$\widehat{state}(t+3)$
$Loss\ = |state(t+3) - \widehat{state}(t+3)|$

Fig. 5 Workflow of the MPUGAT's hybrid method, combining the multi-patch compartment model and neural network. Solid lines represent data input, dashed line represents model training, and gray squares indicate preprocessed data

state(t) values that are nearly identical to the actual values. Conversely, for larger values of h, the focus shifts from inferring parameters to addressing the prediction problem.

Case study

Data

This study employs the proposed model to estimate the dynamic transmission matrix of COVID-19 in South Korea. In this study, the term "city" is used as a generalized representation of geographic regions. For the case study, we use data aggregated at the provincial level, this is conceptually equivalent to the "city" level in the context of our model. We utilize daily confirmed case data for 16 provinces, obtained from the Korea Disease Control and Prevention Agency (KDCA) [50]. The data spans the period from March 1, 2020, to November 23, 2021, preceding the emergence of the Omicron variant. It is assumed that individuals undergo testing within their respective provinces of residence. Due to data limitations, Sejong, a city with limited movement data, is incorporated into the calculations for Chungnam province. Utilizing Eq. (5), we compute the daily number of individuals in each compartment (Susceptible, Exposed, Infectious, and Quarantined). Epidemiological parameters are set as follows: the latent period (α) is fixed at 6.5 days, consistent with findings from Alene et al. [51], and the duration from infection to case confirmation (q) is set to 4 days, based on the analysis by [52].

The model requires two data inputs: time series data for each province and static inter-provincial movement data. Notably, the time series data need not be directly related to infectious disease transmission. In this case study, we utilize daily confirmed COVID-19 cases [50] and Facebook movement data [53] as proxies for provincial dynamics. Specifically, we employ a population weighted sum of the "Change in Movement" and "Stay Put" metrics from the Facebook movement data, capturing shifts in mobility before and after the onset of the COVID-19 pandemic. These metrics allow us to quantify changes in movement patterns during lockdown periods. 'Change in Movement' measures the overall reduction in mobility compared to the two-year pre-pandemic baseline, while 'Stay Put' represents the percentage of people remaining within confined areas. Factors such as unregistered residents, visitors, or temporary relocations during the pandemic could cause the measured 'Stay Put' population to exceed the official registered population count, resulting in percentages above 100%. This data source has been widely adopted in studies investigating human mobility patterns during pandemics [43, 54, 55]. Daily confirmed cases serve as a prominent indicator of epidemic dynamics within each city, with variations in these dynamics intrinsically reflecting underlying differences in intercity mobility. From the daily confirmed cases, we calculated the *state* for each city. These state values, along with the "Change in Movement" and "Stay Put" metrics from Facebook movement data, were used as input features for the model.

To capture the static inter-provincial movement patterns, we utilize traffic Origin-Destination (OD) data procured from the Korea Transport Institute [56]. This dataset, collected in 2018, encompasses major modes of transportation between cities, with origins and destinations determined based on stays exceeding 25 minutes. The deliberate selection of pre-pandemic traffic data enables us to assess the model's adaptive capacity to update and refine its estimates by incorporating information from the COVID-19 period. Based on the aforementioned time series data for each city, we aim to infer

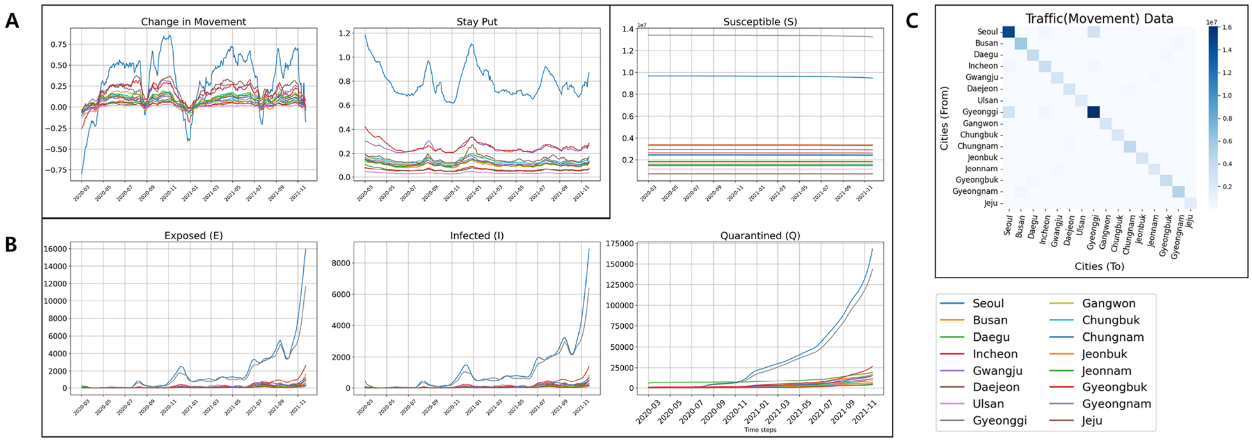


Fig. 6 Model Inputs **A** Facebook movement data: Stay ratio, Change in Movement (Time × City) **B** State data for 16 provinces (Time × City) **C** Traffic data (rows : origin, columns : destination) (City × City)

the dynamic transmission matrix by updating this static inter-city movement data. This approach allows us to investigate how mobility patterns during the pandemic influence disease transmission dynamics. Figure 6 provides a visual representation of the data.

Detailed demographic and geographic information about the provinces in Korea as of November 2021 is shown in Fig. 7 and Table 2.

Korea exhibits substantial spatial heterogeneity across its 16 administrative regions. Notably, the Seoul metropolitan area, encompassing Seoul and Gyeonggi, stands out due to its high cumulative infection count, as it accommodates more than half of the country's population. As shown in Fig. 6A, Seoul and Gyeonggi (represented by the blue and gray lines) follow a similar infection trajectory. This correlation is further corroborated by Fig. 6C, which illustrates the high intercity traffic volume between Seoul and Gyeonggi, excluding intra-city movement (diagonal elements). This high intercity mobility attributed to Gyeonggi's geographical position encircling Seoul, necessitating travel through Gyeonggi when moving from Seoul to other cities. Figure 6C also reveals that intra-city mobility (diagonal elements) generally exceeds intercity traffic. Since Fig. 6C presents raw (non-normalized) values in relation to population, the observed traffic patterns generally correlate with city populations. However, tourist destinations such as Jeju Island exhibit disproportionately high intra-city mobility relative to their population size.

Setting

To preprocess the data, a 14-day moving average filter is applied to mitigate short-term fluctuations and accentuate longer-term trends. This 14-day duration aligns with the conservative incubation period of COVID-19, capturing the potential delay between infection and symptom onset. The inputs were normalized using min-max scaling, which rescales each feature to a range

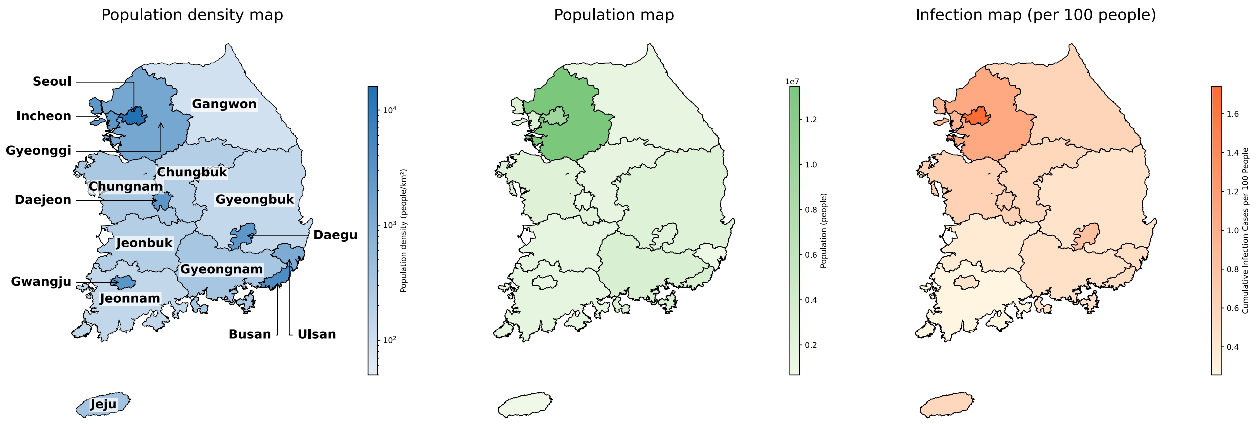


Fig. 7 Population density, Population, and Cumulative Infection Rate data

Table 2 Demographic and infection statistics for cities in Korea (source: [57])

City	Population (1000)	Area(km^2)	Population Density	Cumulative Infections	Infection Rate (%)
Seoul	9,618	605	15,891	168,281	1.75
Busan	3,356	770	4,358	17,505	0.52
Daegu	2,414	883	2,733	19,648	0.81
Incheon	2,951	1,065	2,770	26,791	0.91
Gwangju	1,480	501	2,952	6,530	0.44
Daejeon	1,492	540	2,764	9,362	0.63
Ulsan	1,139	1,062	1,072	5,791	0.51
Gyeonggi	13,452	10,199	1,319	143,644	1.07
Gangwon	1,519	16,878	90	9,396	0.62
Chungbuk	1,631	7,414	220	9,288	0.57
Chungnam	2,525	8,706	290	15,527	0.61
Jeonbuk	1,806	8,062	224	6,802	0.38
Jeonnam	1,793	12,366	145	4,727	0.26
Gyeongbuk	2,652	19,079	139	12,014	0.45
Gyeongnam	3,340	10,536	317	15,862	0.47
Jeju	669	1,853	361	3,905	0.58

between 0 and 1. The movement matrix was reformed to be symmetric and scaled by the learned deep learning parameter b. Subsequently, a rolling 15-day window approach is employed to construct the training dataset. While this 15-day window size is empirically supported, alternative durations could be explored in future research.

- $X(t-16), X(t-15), \ldots, X(t-2) \rightarrow N(t-1)$
- $X(t-15), X(t-14), \ldots, X(t-1) \rightarrow N(t)$

The model was trained using the Adam optimizer with the mean absolute error (MAE) as the loss function, defined as follows:

$$\mathrm{MAE} = \frac{1}{C \cdot T} \sum_{i=1}^{C} \sum_{t=1}^{T} \left| state_i(t+h) - st\hat{a}te_i(t+h) \right| \tag{6}$$

where $state_i$ represents the true values of number of each compartment, and $st\hat{a}te_i$ denotes the corresponding predicted values for city i. To enhance the model's generalization performance, two strategies are employed: hyperparameter tuning and early stopping. Hyperparameter tuning and early stopping were performed on the validation dataset using Bayesian Optimization, as detailed in Table 3. The dataset is divided into training, validation, and test sets, with 80% (481 data points) used for training, 10% (60 data points) for validation, and

Table 3 Hyperparameter tuning and early stopping settings

Hyperparameter space		
lr	[0.0001, 0.01]	Learning rate for model optimization
hidden_unit	[2, 5]	Number of hidden units in the LSTM layers
n_head	[1, 2]	Number of attention heads in the GAT model
lstm_layer	[1, 2]	Number of stacked LSTM layers
Bayesian Optimization Settings (seed : 0)		
Utility Function	EI	Expected Improvement
ξ	0.01	Controls trade-off; lower values favor exploitation
α	$1e^{-4}$	Adds Gaussian noise to avoid overfitting in GP
init_points	10	Number of randomly sampled initial points
n_iter	20	Number of iterations for optimization
Early Stopping Settings		
Patience	10 epochs	Stops if val loss does not decrease for 10 epochs

10% (61 data points) for testing. Additionally, the dataset is divided into sequential mini-batches of 30 data points. According to Algorithm 1, the maximum number of parameters that need to be trained in the deep learning model is 552, which is less than the number of data points in the model.

While previous studies have estimated the spatial transmission matrices by assigning weights based on traffic data or by assuming travel patterns proportional to population size [21, 24, 42, 58], we derive a dynamic transmission matrix $T_{\mathrm{MPUGAT}}(t)$, directly from the proposed MPUGAT model. To validate the efficacy of this transmission matrix, we compare it with the four models listed in Table 4. The first model, the Simple Hybrid Model (SHM), predefines a movement matrix as a contact matrix and estimates the transmission rate (β). In other words, the contact matrix is not dynamically learned but is fixed by applying a symmetric transformation to movement values. Only the transmission rate is inferred using the same spatio-temporal deep learning model as MPUGAT. While SHM utilizes a predefined static contact matrix, MPUGAT dynamically infers the contact matrix from a static movement matrix. Since MPUGAT estimates a greater number of epidemic parameters than SHM, we introduced MPGAT, which employs a static contact matrix derived by averaging the dynamic contact matrices inferred from MPUGAT. This ensures that both SHM and MPGAT infer $\beta(t)$ using the same methodology with predefined contact matrices. All three models can infer dynamic transmission matrix values.

$$\begin{gathered}\tilde{T}_{SHM}(t) = \frac{\tilde{C}_{SHM}}{N} \times \tilde{\beta}_{SHM}(t) = \frac{M}{N} \times \tilde{\beta}_{SHM}(t) \\ \bar{T}_{MPGAT}(t) = \frac{\bar{C}_{MPGAT}}{N} \times \bar{\beta}_{MPGAT}(t) = \frac{\mathrm{Mean}(C_{MPUGAT})}{N} \times \bar{\beta}_{MPGAT}(t) \\ T_{MPUGAT}(t) = \frac{C_{MPUGAT}(t)}{N} \times \beta_{MPGAT}(t)\end{gathered} \tag{7}$$

For $\beta(t)$, it is inferred differently depending on the type of contact matrix, C, to align the dynamics with the predefined C values. Additionally, we include a compartment model that simply uses values from h days ago and a deep learning model as baselines. The deep learning models share the same LSTM architecture as MPUGAT but incorporate an additional MLP to predict *state*(t) for 16 cities, considering only temporal correlations. Furthermore, in deep learning models, min-max normalization is applied to both the input and output, whereas other models normalize only the input.

Results

To quantitatively assess the validity of the estimated parameters, we compare the fitting accuracy using the three types of inferred C and $\beta(t)$ values and deep learning, compartment model. Bayesian optimization is employed to identify the best-performing hyperparameters, and the models are tested with over 30 different seeds, from seed 1 to 30 to evaluate stability. The fitting results for $I(t)$ from the *state*(t) forward 3-day predictions made by the MPUGAT model are illustrated and summarized in Fig. 8.

Despite the differences in dynamics and sizes of $I(t)$ across various cities, it is visually evident that the model fits well. As mentioned in the Training Algorithm section, our model employs a hybrid approach; instead of directly predicting the incidence numbers for each city, as typically done in deep learning models, we estimate epidemic parameters for a multi-patch compartment model. Consequently, even in areas with low incidence numbers, our model can accurately fit the values without any issues. This can be confirmed in Fig. 9 and Table 5, which compare the five models in terms of both computational efficiency and prediction accuracy. The model computational time refers to the total time required to run the training set until the predefined stopping criteria (e.g., epoch stop condition) are met. As shown, all three hybrid models-MPUGAT, MPGAT, and SHM-outperform the other models, including the deep learning and compartment models. The deep learning model performs the worst in predicting incidence across 16 cities in 4 states. While a larger model could improve results, we ensured a fair comparison by using the same LSTM structure as MPUGAT.

In this research, the primary objective is to infer reliable epidemic parameters from the given data rather

Table 4 Models used for comparison and their descriptions

Model name	Model description	Model output	Parameter inference
MPUGAT	$T_{MPUGAT}(t) = C_{MPUGAT}(t) \times \beta_{MPUGAT}(t)$	$\beta(t), C(t)$	O
Simple Hybrid Model (SHM)	$\tilde{T}_{SHM}(t) = \tilde{C}_{SHM} \times \tilde{\beta}_{SHM}(t)$, where $\tilde{C}_{SHM} = M$ (Movement matrix)	$\tilde{\beta}(t)$	O
MPGAT	$\bar{T}_{MPGAT}(t) = \bar{C}_{MPGAT} \times \bar{\beta}(t)$, where $\bar{C}_{MPGAT} = \mathrm{Mean}(C_{MPUGAT}(t))$	$\bar{\beta}(t)$	O
Deep Learning Model	$\hat{state}(t+h)$ predicted using LSTM networks	$\hat{state}(t)$	X
Compartment Model	$\hat{state}(t+h) = state(t)$, delay model	None	X

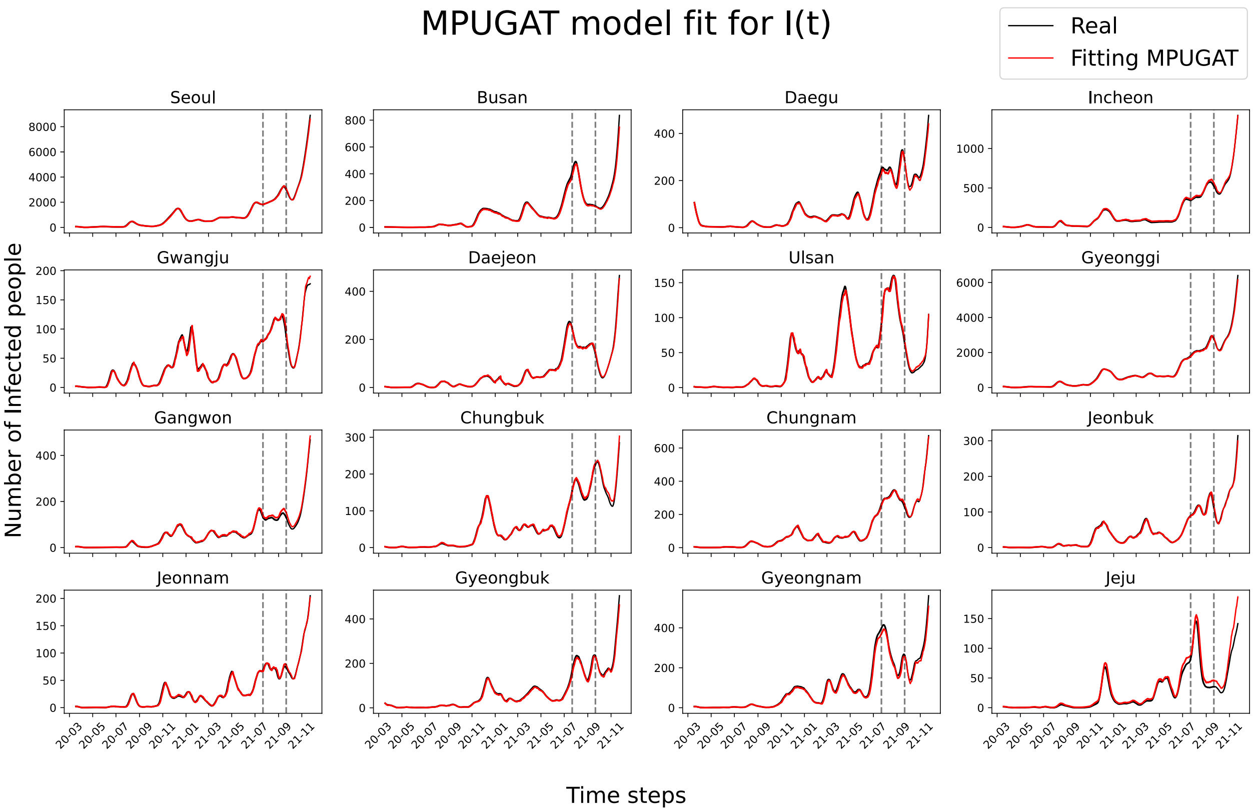


Fig. 8 Fitting results. We conducted 30 iterations to compute the 90% confidence interval for the fitting results. The gray line divides the train, test, and validation datasets

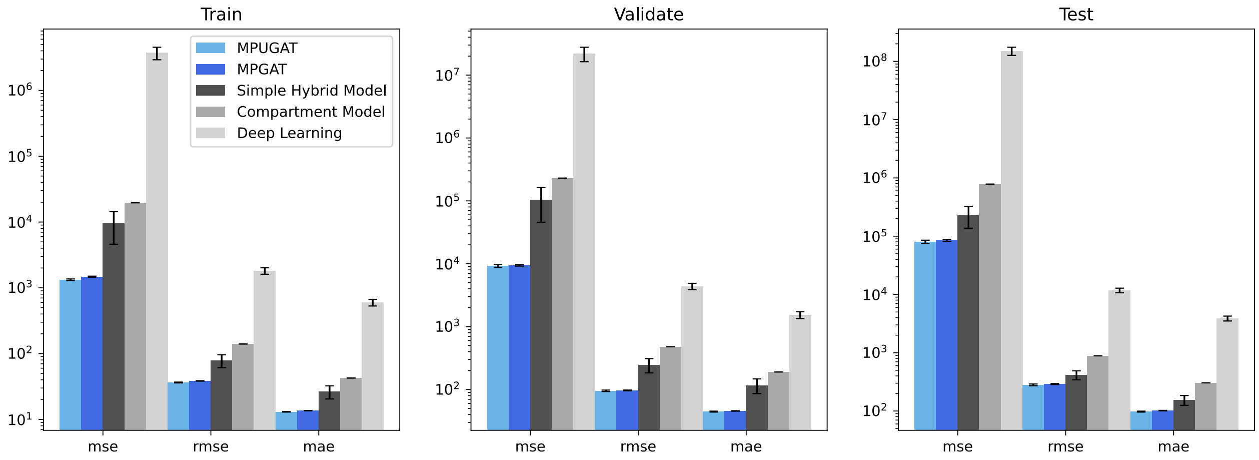


Fig. 9 Comparison of fitting accuracy across models with different transmission matrices. The figure displays the three error metrics-Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE)-as mean values and standard deviations calculated over 30 iterations. The gray bars represent the baseline errors

than to build a prediction model with strong generalization performance. To achieve this, we selected the parameter with optimized performance on training and validation datasets. It is worth noting that test dataset performance, which measures generalization on unseen data, is considered a secondary measure. In our experiments, MPUGAT and MPGAT demonstrated superior performance. However, the difference was not substantial. This highlights the challenge of parameter unidentifiability, as both models describe epidemic dynamics

Table 5 Comparison of model performance

Model	Mode	MSE		RMSE		MAE	
		Mean	Std	Mean	Std	Mean	Std
MPUGAT	Train	1325.88	123.98	36.37	1.69	13.04	0.37
	Validate	9216.55	1674.45	95.66	8.22	44.62	2.50
	Test	80052.10	17540.79	281.27	31.13	97.65	5.91
Hyperparameters: *lr*: 0.009263, *hidden_unit*: 4, *n_head*: 1, *num_lstm_layer*: 1							
Computational Time(s): Training Time: 24.62 ± 7.78 (mean ± std)							
MPGAT	Train	1477.28	82.80	38.42	1.09	13.61	0.17
	Validate	9427.09	792.22	97.02	3.93	45.68	1.42
	Test	84676.34	10266.65	290.43	18.41	101.69	2.99
Hyperparameters: *lr*: 0.003896, *hidden_unit*: 5, *n_head*: 2, *num_lstm_layer*: 2							
Computational Time(s): Training Time: 7.10 ± 2.08							
SHM	Train	9471.56	16273.11	78.63	58.32	26.43	19.91
	Validate	104031.30	194508.10	247.37	210.52	117.13	102.53
	Test	231090.50	314476.20	416.54	244.07	154.81	99.32
Hyperparameters: *lr*: 0.000269, *hidden_unit*: 2, *n_head*: 2, *num_lstm_layer*: 1							
Computational Time(s): Training Time: 68.58 ± 20.25							
Deep Learning	Train	3727870.00	2677651.00	1811.98	678.18	598.59	231.01
	Validate	21988539.00	18955683.00	4379.90	1703.46	1533.61	636.69
	Test	150557345.29	80179206.00	11757.15	3570.98	3879.69	1247.46
Hyperparameters: *lr*: 0.00717, *hidden_unit*: 4, *num_lstm_layer*: 1							
Computational Time(s): Training Time: 6.99 ± 2.84							
Compartment	Train	19626.11	-	140.09	-	42.56	-
	Validate	229660.90	-	479.23	-	190.62	-
	Test	782140.70	-	884.39	-	304.11	-

Hardware Specifications: Intel® Core™ i7-1165G7 (2.80GHz), 16GB RAM, Intel® Iris™ Xe Graphics

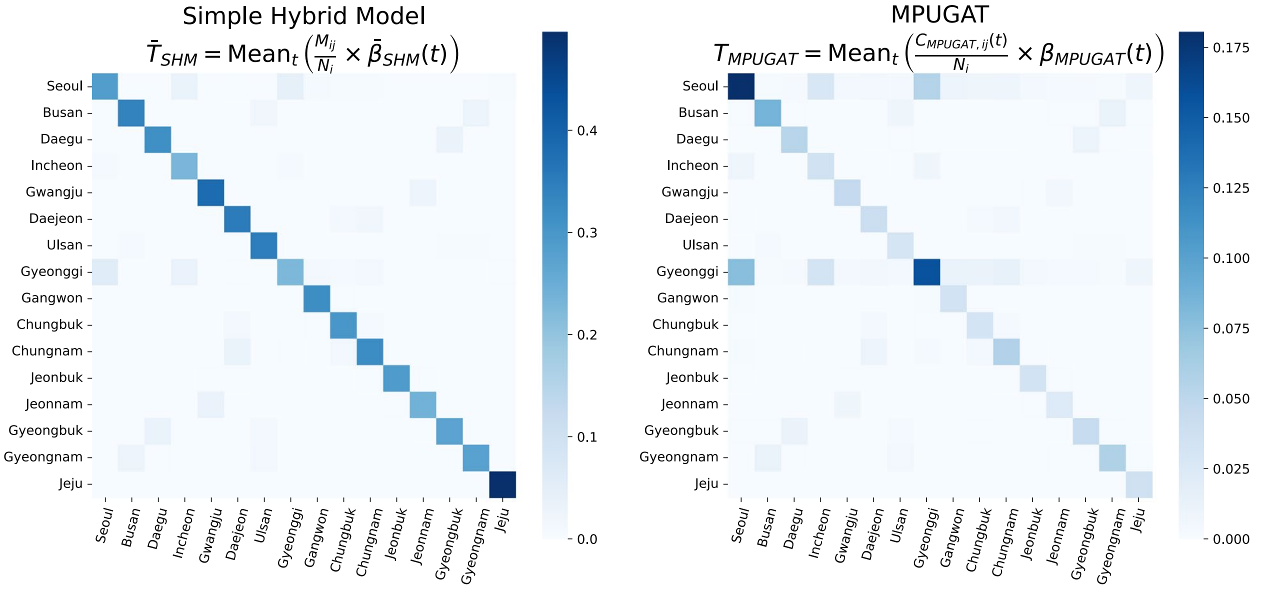


Fig. 10 Comparison between the average values of the transmission matrixs over the entire period obtained from MPUGAT and SHM

despite estimating different parameters. Therefore, our subsequent analysis focuses on qualitative results. As illustrated in Fig. 10, the comparison of transmission matrices derived from SHM and MPUGAT highlights key differences in how these models represent disease transmission patterns.

The transmission matrix T is normalized by the row, N_i for computational convenience in the model. This means that T_{ij} can be interpreted as the probability that a resident of city j is infected by one infected resident of city i. The first distinct aspect is the diagonal elements of the transmission matrices, which represent the probability of infection due to contact among residents of the same city. Since intra-city contact and movement are higher, the diagonal elements have relatively larger values. In Fig. 10, $\bar{T}_{\mathrm{SHM}}$ shows uniformly strong diagonal values, whereas T_{MPUGAT} on the right demonstrates more diverse patterns. Specifically, Jeju Island, a major tourist destination in South Korea, experiences overestimated movement per resident due to external tourists. Jeju residents have a lower population density and more geographically separated urban centers. Consequently, T_{SHM} overestimates the within-resident transmission rate for Jeju while underestimating the rate in the Seoul metropolitan area based solely on movement. In contrast, T_{MPUGAT} captures the transmission inferring lower diagonal values for Jeju and estimating higher values for Seoul, Gyeonggi, and Busan, which are regions with extremely high population densities and strong interconnections between city centers.

The second aspect is the non-diagonal elements of these matrices. As mentioned above, since Seoul and Gyeonggi account for more than half of the total population, we can observe that infected individuals from these regions have a significant impact on other cities. Additionally, Seoul and Gyeonggi have strong influence on other proximity areas (i.e., Incheon, Gangwon, Chungbuk). This demonstrates that MPUGAT considers geographical characteristics in inferring transmission matrices. This pattern is also observed in other similarly-related areas such as Busan-Gyeongnam and Gyeongbuk-Daegu. Exceptionally, Jeju is highly influenced by Seoul even though they are geographically separated, likely due to the air route between them being one of the busiest air routes in the world. In the Gyeonggi-Seoul relationship, a Gyeonggi resident has a stronger influence per capita on Seoul residents than vice versa, reflecting the concentration of workplaces in Seoul while Gyeonggi serves more as a residential area.

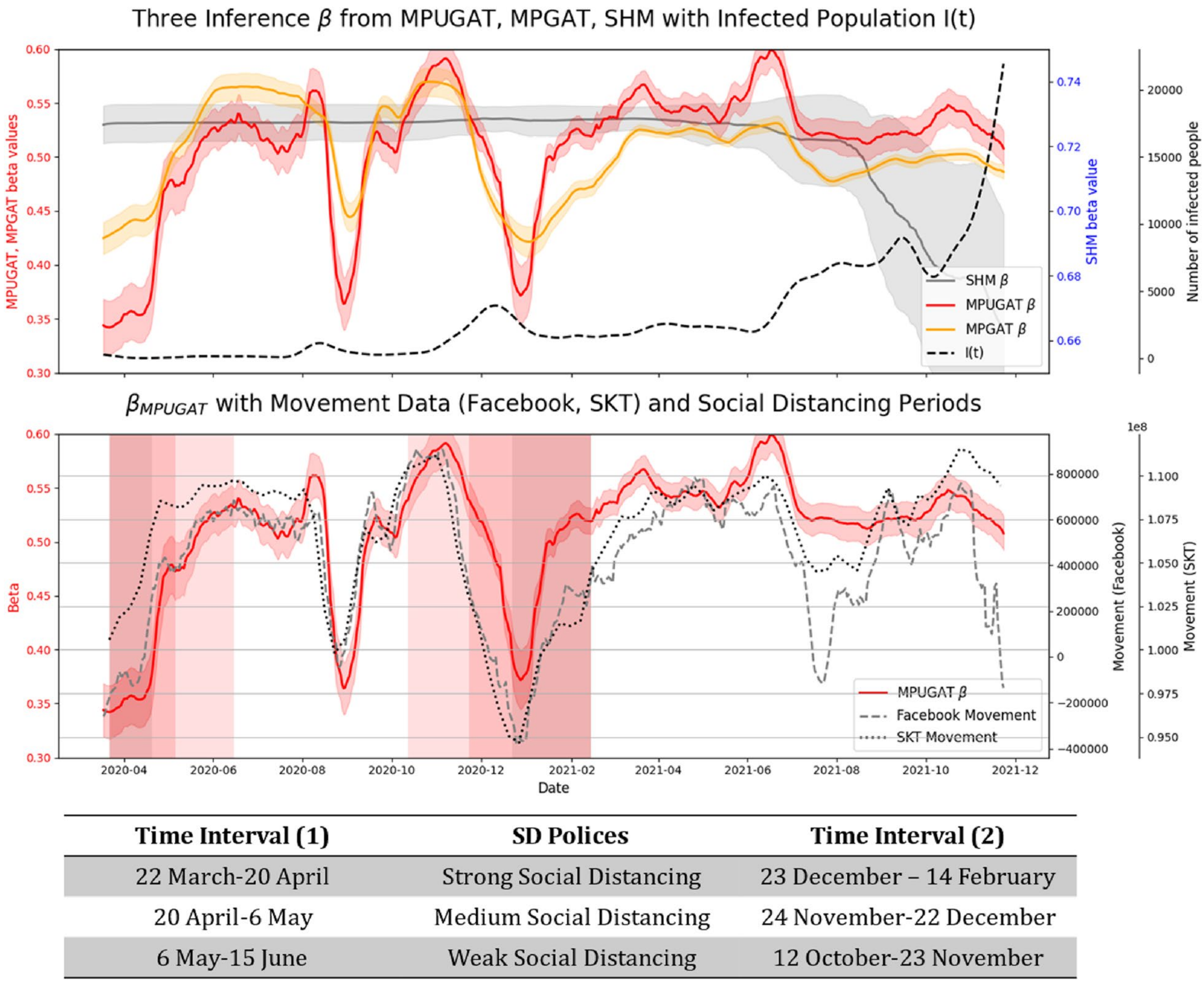
The following examines how the transmission matrix changes over time. We compare the β values inferred from the contact matrix C (or $C(t)$) by the three models: MPUGAT, MPGAT, and SHM. β represents the probability of infection per contact, applied uniformly nationwide, as shown in Fig. 11.

The inferred β_{MPUGAT} and $\tilde{\beta}_{\mathrm{MPGAT}}$ show similar trends. And when compared with the nationwide infected cases I(t), We observe that the β values increase until the $I(t)$ peak is reached, followed by a brief decrease. In contrast, SHM appears to be difficult in deriving meaningful interpretations from its inferred β values. Despite similar fitting accuracy, the epidemic parameters inferred by the two models differ significantly.

To further investigate the impact of external factors, we compare the inferred $\beta(t)$ values with the nationwide social distancing intensity. Following the methodology of [60, 61], we examine the relative changes in social distancing intensity during two specific periods, each exhibiting three levels of intensity. Although social distancing policies were implemented at different times, our focus is on the periods analyzed in the cited studies. Our analysis confirms that $\beta(t)$ values are lower during periods with stricter social distancing measures and higher during periods with less stringent measures, as shown in Fig. 11. However, an interesting observation arises between December 23, 2020, and February 14, 2021. During this period, although social distancing was relatively strong, the $\beta(t)$ values decreased initially and then increased. This phenomenon can be explained by the nationwide movement data, which shows that, despite strict social distancing, the actual movement intensity did not decrease significantly during this time. Furthermore, our analysis of the inferred $\beta(t)$ values suggests that, among the seven input data sources, the contact-based infection probability is most strongly correlated with the population-weighted sum of the change in movement data.

We analyzed the MPUGAT transmission matrix during the three waves of the COVID-19 pandemic in South Korea, as defined in [62]. We visualized the proportion of infections resulting from contact between residents of the same city (without considering the location of the infection) by decomposing the transmission matrix into diagonal and non-diagonal components. The figure illustrates the percentage of such infections within the total infections in residents of each city. The results were visualized in Fig. 12.

Different patterns were observed across the three waves. According to [62], the first wave formed two clusters: Gyeongnam, Gyeongbuk, and Daegu in one, and Seoul and Gyeonggi in another, each with distinct outbreak origins. Our map highlights high intra-resident transmission in Seoul, Gyeonggi, Daegu, Gyeongbuk, and Busan. During the second wave, major infections emerged in Seoul and Gyeonggi, rapidly spreading to Busan. Our visualization supports this, showing decreased intra-resident transmission in Daegu, while Seoul, Gyeonggi,



Time Interval (1)	SD Polices	Time Interval (2)
22 March-20 April	Strong Social Distancing	23 December – 14 February
20 April-6 May	Medium Social Distancing	24 November-22 December
6 May-15 June	Weak Social Distancing	12 October-23 November

Fig. 11 (Upper figure) Comparison of β values from three models (MPUGAT, MPGAT, SHM) with nationwide confirmed cases, $I(t)$. (Lower figure) MPUGAT β with social distancing intensity in South Korea and movement data. The social distancing intensity during Time Interval 1 and Time Interval 2 was assessed using different criteria for each period. The movement data come from two sources: (1) Facebook Movement, nationwide aggregated Change in Movement from [53], and (2) SKT movement data, which collects movement counts from SKT users (a major telecommunications company in Korea) from [59]

and Busan maintained high transmission. The third wave saw nationwide virus spread, reflected in our map by increased intra-resident transmission across many regions. Additionally, the Daejeon-Chungnam and Jeonnam-Gwangju pairs exhibited distinct transmission patterns that evolved independently across the three waves, consistent with prior findings.

Conclusion and discussion

In this study, we presented MPUGAT, a novel hybrid framework that integrates a multi-patch compartmental model with a spatio-temporal deep learning model to dynamically estimate the transmission matrix of infectious diseases. We decomposed the transmission matrix into movement data and resident ratio matrices, following the mass action principle. And then, By utilizing readily available data through the attention mechanism, MPUGAT generates a dynamic transmission matrix. While GAT is used for predicting the dynamics of graph-structured data, this study leverages the attention mechanism to expand the dimensionality of data. This approach addresses the limitations related to data availability and the identifiability problem often encountered in prior modeling studies. Moreover, this framework enables the integration of diverse data types beyond traditional epidemiological metrics, enhancing model flexibility without incurring high-dimensional complexity. In light of the growing volume of available data, MPUGAT's

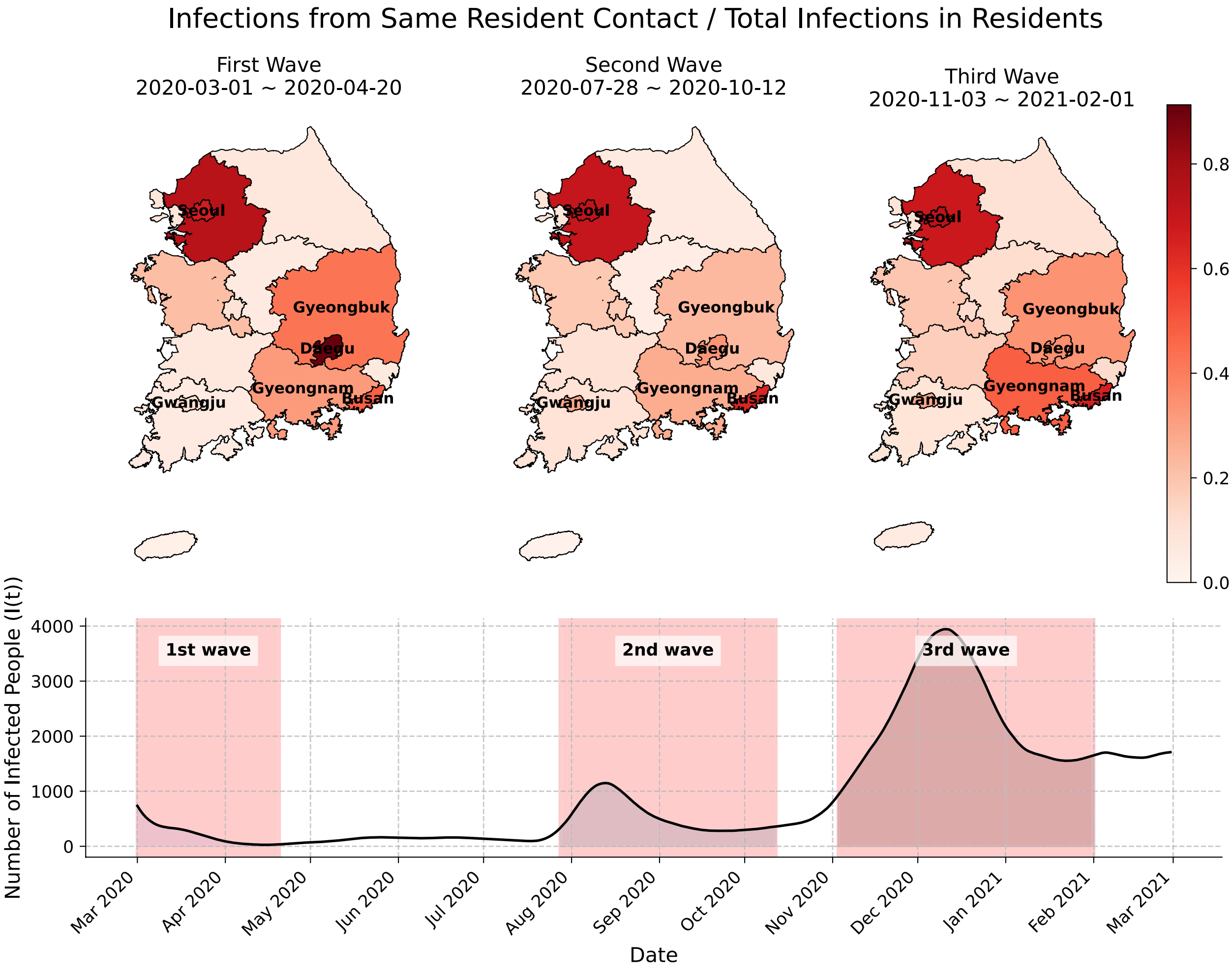


Fig. 12 Proportion of infections from same-resident contact relative to total infections in residents across three COVID-19 waves in South Korea. (Upper panel): Choropleth maps showing the ratio of within-resident transmission for each province during the first wave (March 1 - April 20, 2020), second wave (July 28 - October 12, 2020), and third wave (November 3, 2020 - February 1, 2021). Darker red indicates higher proportions of same-resident transmission. (Lower panel): Time series showing the number of infected people nationwide with highlighted wave periods

adaptability in incorporating multiple time-series sources suggests numerous possibilities for future research. Future research could explore incorporating data sources such as social media trends, news reports, and economic indicators to further enrich MPUGAT's predictive capacity. In our case study, we employed actual movement data of each city in South Korea, enabling MPUGAT to provide a more accurate and dynamic representation of disease spread. While MPUGAT currently focuses more on analysis than real-time prediction, it could be extended to applications beyond epidemic modeling to include real-time processing of traffic information and other domains. In such extensions, computational efficiency could be further enhanced by implementing tensor decomposition-based spatial-temporal information compression techniques as proposed by [63] or parallel deep reinforcement learning computing structures [64].

Our case study using COVID-19 data from South Korea demonstrated the effectiveness of MPUGAT in capturing temporal variations in transmission dynamics and providing a more nuanced understanding of disease spread compared to traditional methods. The model successfully captured the expected high contact rates within densely populated metropolitan areas and identified changes in the infection characteristics of cities during different epidemic periods. Furthermore, comparing the estimated transmission rates from MPUGAT with actual policies, such as social distancing, confirms that stricter measures correspond to lower transmission rates. Likewise, nationwide mobility data exhibits a similar relationship with β.

Accurate inference of the transmission matrix, reflecting city characteristics, is crucial for developing tailored public health policies. The transmission matrices derived from the MPUGAT and SHM models exhibit significant differences. For instance, applying uniform guidelines to all cities may be less effective than implementing targeted approaches based on city-specific contact patterns. In this context, the two transmission matrices highlight different areas of concern-MPUGAT emphasizes Seoul and Gyeonggi, while SHM focuses on Jeju. Moreover, the β values inferred by the two models show substantial differences in their interpretability, leading to significant discrepancies in epidemic scenario analyses. While SHM achieves good epidemic fitting, its transmission matrix fails to produce meaningful insights. In contrast, the β values from MPUGAT and MPGAT meaningfully reflect social distancing measures, nationwide mobility trends, and the number of confirmed cases. Furthermore, the epidemic parameters provided by MPUGAT in Fig. 12 indicate that the cities requiring careful attention vary over time. This adaptability enables timely public health interventions and the efficient allocation of limited healthcare resources. By capturing temporal variations in transmission patterns, MPUGAT can support policymakers in implementing targeted interventions, optimizing resource distribution, and assessing the effectiveness of different strategies.

While MPUGAT offers a promising approach for infectious disease modeling, this study has certain limitations. First, the accuracy of MPUGAT depends on the availability and quality of data. Inaccurate or incomplete data may lead to biased estimations and misleading conclusions. In particular, to effectively utilize the transmission matrices in public health applications, proper validation is essential. Specifically, since there is no predefined ground truth for transmission matrices, validating the inferred values requires additional external epidemiological analysis. Second, inherent variability in deep learning models, stemming from their probabilistic nature, can influence results. When inferring high-dimensional parameters, the model identifies the parameter set that maximizes the probability given the data, which may not always yield consistent outcomes. Thus, conducting multiple model iterations, as demonstrated in this study, is essential to ensure the robustness and validity of the results. Third, while our case study focused on COVID-19 in South Korea, the generalizability of MPUGAT to other infectious diseases and geographical settings needs further investigation. Furthermore, the SEIQ model used in this study is a simplification of the complex dynamics of infectious disease transmission. More complex compartmental models could be incorporated to account for factors such as vaccination, asymptomatic infections, and different levels of disease severity.

Despite these limitations, MPUGAT provides a valuable tool for understanding hidden epidemiological patterns and supporting public health decision-making by estimating more reliable transmission matrices. Its ability to dynamically estimate transmission parameters and incorporate diverse data sources, while maintaining a small model size independent of the number of cities analyzed, enhances its potential for understanding and predicting disease spread. By addressing the limitations identified in this study and further refining the model, we can enhance its accuracy, generalizability, and applicability to a wider range of public health challenges.

Acknowledgements
Not applicable.

Authors' contributions
ML: formal analysis, visualization, data acquisition, and methodology. ML and CHL: conceptualization, and writing—original draft. ML, HC, and CHL: validation, and writing—review & editing. CHL: funding acquisition, and supervision. All authors read and approved the final manuscript.

Funding
This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2022R1F1A1064487) and the BK21 Program (Next Generation Education Program for Mathematical Sciences, 4299990414089) funded by the Ministry of Education(MOE, Korea) and National Research Foundation of Korea (NRF).

Data availability
The datasets generated during and/or analysed during the current study are available in the [MPUGAT] repository, https://github.com/hellominji13/MPUGAT.

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Competing interests
The authors declare no competing interests.

Received: 4 December 2024 Accepted: 6 May 2025
Published online: 22 May 2025

References

1. Jones BA, Betson M, Pfeiffer DU. Eco-social processes influencing infectious disease emergence and spread. Parasitology. 2017;144(1):26–36. https://doi.org/10.1017/S0031182016001414.
2. Desai AN, Kraemer MUG, Bhatia S, Cori A, Nouvellet P, Herringer M, et al. Real-time Epidemic Forecasting: Challenges and Opportunities. Health Secur. 2019;17(4):268–75. https://doi.org/10.1089/hs.2019.0022.
3. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proc R Soc Lond Ser A Containing Pap Math Phys Character. 1927;115(772):700–21.
4. Isham V, Medley G. Models for infectious human diseases: their structure and relation to data, vol. 6. Cambridge: Cambridge University Press; 1996.

5. Segbroeck SV, Santos FC, Pacheco JM. Adaptive Contact Networks Change Effective Disease Infectiousness and Dynamics. PLoS Comput Biol. 2010;6(8):e1000895.
6. Rohani P, Zhong X, King AA. Contact Network Structure Explains the Changing Epidemiology of Pertussis. Science. 2010;330(6006):982–5.
7. Keeling MJ, Eames KTD. Networks and epidemic models. J R Soc Interf. 2005;2(4):295–307. https://doi.org/10.1098/rsif.2005.0051.
8. Andersson H. Epidemics in a population with social structures. Math Biosci. 1997;140(2):79–84. https://doi.org/10.1016/S0025-5564(96)00129-0.
9. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med. 2008;5(3):e74.
10. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases. PLoS Med. 2008;5(3):1. https://doi.org/10.1371/journal.pmed.0050074.
11. Keeling MJ. The effects of local spatial structure on epidemiological invasions. Proc R Soc Lond Ser B Biol Sci. 1999;266(1421):859–67. https://doi.org/10.1098/rspb.1999.0716.
12. Wellenius GA, Vispute SS, Espinosa V, Fabrikant A, Tsai T, Hennessy J, et al. Impacts of social distancing policies on mobility and COVID-19 case growth in the US. Nat Commun. 2021;12:3118.
13. Kim S, Castro MC. Spatiotemporal pattern of COVID-19 and government response in South Korea (as of May 31, 2020). Int J Infect Dis. 2020;98:328–33.
14. Kang D, Choi J, Kim Y, Kwon D. An analysis of the dynamic spatial spread of COVID-19 across South Korea. Sci Rep. 2022;12(1):9364.
15. Ball F, Mollison D, Scalia-Tomba G. Epidemics with two levels of mixing. Ann Appl Probab. 1997;7(1):46–89.
16. Sattenspiel L, Dietz K. A structured epidemic model incorporating geographic mobility among regions. Math Biosci. 1995;128(1–2):71–91.
17. Arino J, van den Driessche P. A multi-city epidemic model. Math Popul Stud. 2003;10(3):175–93.
18. Pujari BS, Shekatkar S. Multi-city modeling of epidemics using spatial networks: Application to 2019-nCov (COVID-19) coronavirus in India. medRxiv [preprint]. 2020. https://doi.org/10.1101/2020.03.13.20035386.
19. Fumanelli L, Ajelli M, Manfredi P, Vespignani A, Merler S. Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. PLOS Comput Biol. 2012;8(9):e1002673.
20. Kim JE, Choi H, Lee M, Lee CH. The effect of shortening the quarantine period and lifting the indoor mask mandate on the spread of COVID-19: a mathematical modeling approach. Front Public Health. 2023;11:1166528.
21. Wei Y, Wang J, Song W, Xiu C, Ma L, Pei T. Spread of COVID-19 in China: analysis from a city-based epidemic and mobility model. Cities. 2021;110:103010.
22. Ciofi degli Atti ML, Merler S, Rizzo C, Ajelli M, Massari M, Manfredi P, et al. Mitigation measures for pandemic influenza in Italy: an individual based model considering different scenarios. PLoS ONE. 2008;3(3):e1790.
23. Dai D. Racial/ethnic and socioeconomic disparities in urban green space accessibility: Where to intervene? Landsc Urban Plan. 2011;102(4):234–44.
24. Lunelli A, Pugliese A, Rizzo C. Epidemic patch models applied to pandemic influenza: contact matrix, stochasticity, robustness of predictions. Math Biosci. 2009;220(1):24–33.
25. Raissi M, Ramezani N, Seshaiyer P. On parameter estimation approaches for predicting disease transmission through optimization, deep learning and statistical inference methods. Lett Biomath. 2019;6(2):1–26.
26. Ball F, Britton T, House T, Isham V, Mollison D, Pellis L, et al. Seven challenges for metapopulation models of epidemics, including households models. Epidemics. 2015;10:63–7.
27. Volz EM, Meyers L. Susceptible-infected-recovered epidemics in dynamic contact networks. Proc R Soc B Biol Sci. 2007;274:2925–34.
28. Truscott J, Ferguson NM. Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. PLoS Comput Biol. 2012;8(10):e1002699.
29. Tizzoni M, Bajardi P, Decuyper A, Kon Kam King G, Schneider CM, Blondel V, et al. On the use of human mobility proxies for modeling epidemics. PLoS Comput Biol. 2014;10(7):e1003716.
30. Villaverde AF, Barreiro A, Papachristodoulou A. Structural identifiability of dynamic systems biology models. PLoS Comput Biol. 2016;12(10):e1005153.
31. Chowell G. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. Infect Dis Model. 2017;2(3):379–98.
32. Clement JC, Ponnusamy V, Sriharipriya K, Nandakumar R. A survey on mathematical, machine learning and deep learning models for COVID-19 transmission and diagnosis. IEEE Rev Biomed Eng. 2021;15:325–40.
33. Liu Z, Wan G, Prakash BA, Lau MS, Jin W. A review of graph neural networks in epidemic modeling. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery; 2024. p. 6577–87.
34. Yu B, Yin H, Zhu Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. arXiv [preprint] arXiv:1709.04875. 2017. https://arxiv.org/abs/1709.04875.
35. Gao J, Sharma R, Qian C, Glass LM, Spaeder J, Romberg J, et al. STAN: spatio-temporal attention network for pandemic prediction using real-world evidence. J Am Med Inform Assoc. 2021;28(4):733–43.
36. Berkhahn S, Ehrhardt M. A physics-informed neural network to model COVID-19 infection and hospitalization scenarios. Adv Contin Discret Model. 2022;2022(1):61.
37. Kharazmi E, Cai M, Zheng X, Zhang Z, Lin G, Karniadakis GE. Identifiability and predictability of integer-and fractional-order epidemiological models using physics-informed neural networks. Nat Comput Sci. 2021;1(11):744–53.
38. Grimm V, Heinlein A, Klawonn A, Lanser M, Weber J. Estimating the time-dependent contact rate of SIR and SEIR models in mathematical epidemiology using physics-informed neural networks. Electron Trans Numer Anal. 2022;56:1–27.
39. Bousquet A, Conrad WH, Sadat SO, Vardanyan N, Hong Y. Deep learning forecasting using time-varying parameters of the SIRD model for Covid-19. Sci Rep. 2022;12(1):3030.
40. Jo H, Son H, Hwang HJ, Jung SY. Analysis of COVID-19 spread in South Korea using the SIR model with time-dependent parameters and deep learning. medRxiv [preprint]. 2020. https://doi.org/10.1101/2020.04.13.20063412.
41. Liao Z, Lan P, Fan X, Kelly B, Innes A, Liao Z. SIRVD-DL: A COVID-19 deep learning prediction model based on time-dependent SIRVD. Comput Biol Med. 2021;138:104868.
42. Mao J, Han Y, Wang B. MPSTAN: Metapopulation-Based Spatio-Temporal Attention Network for Epidemic Forecasting. Entropy. 2024;26(4):278.
43. Cao Q, Jiang R, Yang C, Fan Z, Song X, Shibasaki R. MepoGNN: Metapopulation epidemic forecasting with graph neural networks. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham:Springer; 2022. p. 453–68.
44. Rahmadani F, Lee H. Hybrid deep learning-based epidemic prediction framework of COVID-19: South Korea case. Appl Sci. 2020;10(23):8539.
45. Wang L, Adiga A, Chen J, Sadilek A, Venkatramanan S, Marathe M. Causalgnn: Causal-based graph neural networks for spatio-temporal epidemic forecasting. In: Proceedings of the AAAI conference on artificial intelligence. Palo Alto: AAAI Press; 2022;36(11):12191–9.
46. Kim D, Oh A. How to find your friendly neighborhood: Graph attention design with self-supervision. arXiv [preprint]. 2022. arXiv:2204.04879. https://arxiv.org/abs/2204.04879.
47. Lenormand M, Picornell M, Cantú-Ros OG, Louail T, Herranz R, Barthelemy M, et al. Comparing and modelling land use organization in cities. R Soc Open Sci. 2015;2(12):150449.
48. Dong X, Morales AJ, Jahani E, Moro E, Lepri B, Bozkaya B, et al. Segregated interactions in urban and online space. EPJ Data Sci. 2020;9(1):20.
49. Sobolevsky S, Sitko I, Des Combes RT, Hawelka B, Arias JM, Ratti C. Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. The case of residents and foreign visitors in Spain. In: 2014 IEEE International Congress on Big Data. Piscataway: IEEE; 2014. p. 136–43.
50. Korea Disease Control and Prevention Agency (KDCA). Provincial-level COVID-19 case data in South Korea. 2020-2022. Online. https://dportal.kdca.go.kr/pot/cv/trend/dmstc/selectMntrgSttus.do. Accessed 18 Nov 2024.
51. Alene M, Yismaw L, Assemie MA, Ketema DB, Gietaneh W, Birhan TY. Serial interval and incubation period of COVID-19: a systematic review and meta-analysis. BMC Infect Dis. 2021;21:1–9.
52. Park MB, Park EY, Lee TS, Lee J. Effect of the period from COVID-19 symptom onset to confirmation on disease duration: quantitative analysis of publicly available patient data. J Med Internet Res. 2021;23(9):e29576.

53. Facebook Data for Good. Facebook Movement Range data. 2020-2022. Online. https://data.humdata.org/dataset/movement-range-maps. Accessed 18 Nov 2024.
54. Shepherd HE, Atherden FS, Chan HMT, Loveridge A, Tatem AJ. Domestic and international mobility trends in the united kingdom during the covid-19 pandemic: An analysis of facebook data. Int J Health Geogr. 2021;20:1–13.
55. Chan J. The geography of social distancing in Canada: Evidence from Facebook. Can Public Policy. 2020;46(S1):S19–28.
56. Korea Transport Institute (KOTI). Traffic OD (Origin-Destination) data. 2020-2022. Online. https://www.koti.re.kr/eng.do. Accessed 16 Aug 2022.
57. Service KSI. Population and Demographic Statistics. 2020. https://www.index.go.kr/unity/potal/main/EachDtlPageDetail.do?idx_cd=1007. Accessed Feb 2025.
58. Bomfim R, Pei S, Shaman J, Yamana T, Makse HA, Andrade JS Jr, et al. Predicting dengue outbreaks at neighbourhood level using human mobility in urban areas. J R Soc Interface. 2020;17(171):20200691.
59. SKT (SK Telecom) and Statistics Korea. SKT Movement Data. 2020-2021. Online. https://data.kostat.go.kr/nowcast/main.do?initId=19. Accessed 04 Mar 2025.
60. Choi Y, Kim JS, Choi H, Lee H, Lee CH. Assessment of social distancing for controlling COVID-19 in Korea: an age-structured modeling approach. Int J Environ Res Public Health. 2020;17(20):7474.
61. Choi Y, Kim JS, Kim JE, Choi H, Lee CH. Vaccination prioritization strategies for COVID-19 in Korea: a mathematical modeling approach. Int J Environ Res Public Health. 2021;18(8):4240.
62. Kim S, Kim M, Lee S, Lee YJ. Discovering spatiotemporal patterns of COVID-19 pandemic in South Korea. Sci Rep. 2021;11(1):24470.
63. Xu Z, Lv Z, Chu B, Li J. A fast spatial-temporal information compression algorithm for online real-time forecasting of traffic flow with complex nonlinear patterns. Chaos Solitons Fractals. 2024;182:114852.
64. Lv Z, Li J, Xu Z, Wang Y, Li H. Parallel computing of spatio-temporal model based on deep reinforcement learning. In: International Conference on Wireless Algorithms, Systems, and Applications. Cham: Springer International Publishing; 2021. p. 391–403.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.