IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 July 2024, accepted 9 August 2024, date of publication 20 August 2024, date of current version 4 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3446619

RESEARCH ARTICLE

A Deep Graph Neural Network Approach for Assessing Origin Destination Traffic Flow Estimates Based on COVID-19 Data

MARIO MUÑOZ-ORGANERO [1], (Senior Member, IEEE), AND VÍCTOR CORCOBA-MAGAÑA [2]

[1]Department of Telematic Engineering, Universidad Carlos III de Madrid, Leganes, 28911 Madrid, Spain
[2]Department of Computer Science, University of Oviedo, 33003 Oviedo, Spain

Corresponding author: Mario Muñoz-Organero (munozm@it.uc3m.es)

This work was supported in part by the Project "Generation of Reliable Synthetic Health Data for Federated Learning in Secure Data Spaces" Research Project (Agencia Estatal de Investigación (AEI)/European Regional Development Fund (ERDF), EU) funded by Ministerio de Ciencia, Innovación y Universidades (MCIN)/AEI/10.13039/501100011033 under Grant PID2022-141045OB-C43 and by "ERDF a Way of Making Europe" through "European Union"; in part by the Project "FLATCITY-APP: Aplicación móvil para FlatCity" funded by Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación MCIN/AEI/10.13039/501100011033 and by European Union "NextGenerationEU/Plan de Recuperación, Transformación y Resiliencia (PRTR)" under Grant PDC2021-121239-C33; and in part by the Project Programa de Universalización de Infraestructuras Digitales para la Cohesión (UNICO) 5G I+D 6G-RIEMANN funded by Spanish Ministry of Economic Affairs and Digital Transformation and European Union-NextGenerationEU.

ABSTRACT Origin-Destination (OD) traffic flow estimations from traffic sensor data play an important role for transportation planning and management. This paper proposes a novel method to compare OD traffic estimated matrices (using data from traffic sensors). The proposed method uses the estimated OD traffic flow values together with COVID-19 incidence data in order to build a sequence of temporal graphs that are fed into a machine learning (ML) model. The ML model uses the input information to estimate/predict one-week ahead COVID-19 incidence data. A tailored Graph Neural Network (GNN) and Long Short-Term Memory (LSTM) model is designed adapted to the input information. The paper evaluates the proposed method with 3 different OD estimation alternatives and compares the accuracy achieved by different configurations of the ML model with a traffic agnostic baseline model. Data from 44 provinces in Spain during 2021 providing daily COVID-19 incidence data and 635 geo-located traffic sensors providing monthly traffic counts are used to evaluate the results. The 3 traffic-aware OD estimation methods were able to outperform the baseline model, achieving model gains up to 136%. The major application of the results of this paper is a novel mechanism to validate prior OD traffic matrices.

INDEX TERMS Intelligent systems for traffic prediction, traffic modeling, traffic flow modeling, machine learning, artificial intelligence, prior OD traffic flow estimation, graph neural network, long short-term memory recurrent neural network, COVID-19 virus propagation.

I. INTRODUCTION

Traffic sensors on roads are able to capture different data sources, ranging from simple vehicle counts (using sensors such as loop detectors) to particular vehicle trajectories based on cameras and Automatic Number Plate Recognition (ANPR) [1]. Traffic flow data collected by traffic sensing devices is crucially important for transportation planning and transportation management [2]. Since the number of traffic sensing devices in road networks is normally limited due to their high installation and maintenance costs [2], alternative techniques have been proposed in previous research studies to provide estimations of traffic flows in non-monitored road links and to estimate origin-destination (OD) traffic matrices that provide an overall view of the traffic demand in an entire network.

Cho et al. classified the different methodologies to estimate OD matrices into survey based, trip distribution models and non-assignment-based models based on traffic counts [3]. Surveys are expensive to carry out and provide only a fixed

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

image in a particular time. Trip distribution models such as gravity models, provide a theoretical framework to estimate traffic flows based on underlaying variables that have an impact on them, such as demographics, traffic generation models, zone attraction models and travel cost functions. Several traffic assignment models have been used in order to map prior calculated OD matrices to traffic flows and traffic sensors at particular road segments have been used to generate iterative methods to improve prior OD matrices based on traffic measured data. Traffic estimation based on iterative methods tend to require a detailed knowledge of the road network and the time required to cross each segment of the network.

Other sensors have also been used to enrich the information obtained from traffic count sensors and provide better estimates for unobserved traffic flows and OD matrices. Wang et al. [2] used License Plate Recognition (LPR) sensors and taxi GPS trajectory data to enhance the information from 68 traffic flow sensors which were able to only measure 1.2% of the entire road network. Liu et al. [4] used Cellphone Location (CL) and License Plate Recognition (LPR) data for estimating traffic flows in a large road network using a multi-source model. Adding more sensors to the road network or requiring the user of the network to provide personal location data introduced new challenges both in terms of costs and data sharing.

Based on traffic data from sensors, a new approach to estimate traffic flows and OD matrices based on the use of machine learning models have been explored over the last years. Machine learning models are able to extract patterns from sensor data in order to provide estimates for un-observed data. The algorithms need to be trained with known data before they can be used for providing estimates for new data. The majority of the machine learning models used to estimate traffic flows and OD matrices use a supervised learning approach for which labeled training data is required. Pamula et al. [5] proposed a deep learning model based on the knowledge of the structure of the road network, origin and destination points of trips, as well as data on traffic intensity on road network sections recorded by video-sensing devices, in order to estimate OD matrices. Training data was synthetically generated using prior OD matrices. Liu et al. [6] used mobile phone data in order to estimate dynamic OD matrices.

In this paper, we propose a novel method to evaluate OD traffic matrix estimations based on traffic count sensors that does not require a complete knowledge of the topology of the network and does not require the knowledge of large training sets of OD traffic matrices for using machine learning models. The proposed method uses a Graph Neural Network (GNN) machine learning model to fit the spatial characteristics of a road network and time series of COVID-19 incidence data to get geo-located temporal sequences that capture the spread of the virus. The output of the GNN (a temporal sequence of processed graphs) is fed into a Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) to extract temporal patterns from the input data. The COVID-19 virus is transmitted from human to human interactions and is therefore affected by OD traffic flows. The proposed machine learning model receives OD traffic estimated matrices as input data and uses a graph neural network to optimally estimate the spread of the COVID-19 virus considering the incubation period and the human mobility estimations. The idea motivating the proposed methodology is that better OD traffic estimations will lead to better results for COVID-19 spread predictions (the machine learning model will be able to generate more accurate predictions when the input data is closer to the un-observed real data). The model will be validated using real data from 44 provinces and 635 traffic sensors in road segments in Spain during the entire 2021. The mutual information between mobility and COVID-19 incidence data for the escalation and de-escalation of new COVD-19 cases in Spain for the data from 2021 is analyzed to support the proposed model.

The major contributions of this paper are:

- *Definition of a graph format to combine COVID-19 incidence data, as node information, and OD traffic flow estimates, as edge information.*
- *Proposal of a novel GNN-LSTM machine learning model designed to extract patterns from a COVID-19-OD-Traffic sequence of graphs.*
- *Proposal of a novel methodology to evaluate OD traffic flow estimations that uses the OD estimates to generate input graphs and evaluates the accuracy achieved by the GNN-LSTM model.*

The primary motivation of this research is to have a method to estimate the accuracy for OD traffic matrices based on traffic sensor information not requiring extensive and expensive training data but using the already COVID-19 gathered data. A secondary motivation is to show that the spatial spread of the COVID-19 virus can be better estimated using traffic data and mobility estimations.

The paper is organized as follows. Section I, this section, provides and introduction and motivation of the study. Section II captures the previous related work to justify the novelty of the proposed approach. Section III describes the proposed method. Section IV captures the description of the datasets used in order to validate the results. The datasets contain monthly updated data from traffic sensors in several roads in Spain and daily COVID-19 reported data per province in Spain. Section V is dedicated to analyze the datasets in order to validate the correlation between human mobility and the COVID-19 spread in Spain in 2021 and to evaluate if road traffic data could be considered a good proxy variable for estimating human mobility. Section VI presents 4 different methods, used in this paper to estimate prior OD matrices. Section VII captures the results of the validation and major conclusions are presented in section VIII.

II. RELATED WORK

Traffic counts based on traffic sensors located in sparse locations of a road network provide not only a direct traffic monitoring system but also a data source for methods to estimate overall traffic flows and OD matrices for the entire network. Timms [7] performed a review of methods published between 1970 and 2000 to estimate OD matrices based on a trade-off of available limited traffic link observational data and models to estimate prior subjective input OD matrices. Different methods have been used to generate prior OD matrices such as [8] and [9] and several allocations models could be used to distribute OD traffic into particular network links. Observed link counts could then be used to assess if the calculated data coincides with the observed data and the OD matrices are adjusted in an iterative process.

In recent years, new models to estimate overall traffic data from sensors have been proposed based on the use of additional sensors that are currently available. Wang et al. [2] proposed a data driven model that expanded the information of traffic flow sensors with License Plate Recognition (LPR) data and taxi GPS trajectory data for estimating traffic flow in large road networks. Sánchez-Cambronero et al. [10] proposed a model for estimating dynamic traffic flows in road networks by using Automatic Number of Plate Recognition (ANPR) sensors and a first-in-first-out (FIFO) hypothesis. ANPR sensors were used to evaluate travel times among different points in the network in order to better estimate traffic flows. Yang et al. [11] used probe vehicles to estimate trajectories in the road network. The estimated trajectory information was used to complement observed link count information. The GPS sensors were used on probe vehicles to estimate their trajectories. The model proposed the use of the information from probe vehicles and link counts to estimate prior OD matrices and a conventional generalized least squares (GLS) framework was used to conduct OD correction using link counts. Cho et al. [3] proposed a method that used Gibbs sampling and a Kalman filter to avoid requiring the use of a prior OD matrix. Wu et al. [12] proposed a four-stage traffic flow prediction method based on the use of a gravity model to estimate prior OD matrices and a Dijkstra algorithm to calculate optimal routes. The authors compared two different assignment methods in a theoretical way. Hualan et al. [13] proposed a method to estimate the OD matrix for the particular case of a non-congested ring expressway by using the information about the topology of the network, the on-ramp and off-ramp traffic, based on some basic assumptions about the routes taken by the drivers.

A different approach to estimate traffic flows and OD matrices from observed sensor data is using machine learning models. Pamula and Żochowska [5] proposed a method for OD (Origin–Destination) matrix prediction based on traffic data using Recurrent Neural Networks (RNN) based on Long Short-Term Memory (LSTM) or autoencoder layers (DLNa — deep learning networks with autoencoders). An iterative method was used for estimating OD matrices over historical data to generate training sequences for the neural network. Liu et al. [4] proposed a machine learning model in order to estimate network flow information based on the use of cellphone location and License Plate Recognition (LPR) data. Afandizadeh et al. [1] also used a deep machine learning model to estimate hourly OD traffic matrices. The model used data extracted from automatic number plate recognition (ANPR) cameras, smart fare cards, loop detectors at intersections, global positioning systems (GPS) of navigation software, socio-economic and demographic characteristics as well as land-use features of zones as inputs. The model was trained based on the knowledge of ground truth OD matrices for the city of Tehran based on activity-based data from 2019.

Traffic information has both spatial and temporal components. The spatial information in a road network can be expressed using a directed graph. Liu et al. [6] proposed a machine learning model-based Graph Convolutional Neural Network (GCN) to estimate OD traffic flows. OD traffic data was captured from mobile phone data. The network topology was divided into city areas and the objective of the machine learning model was to estimate future temporal values for the OD traffic matrices based on measured matrices and using a graph to capture the interconnections among different areas. Two different graphs architectures were used by considering OD flows as the nodes or the edges of the graph. Yu et al. [14] used a Spatio-Temporal Graph Convolutional Network to predict the time dependent behavior of traffic.

Traffic flows (and OD traffic matrices) capture the movements of people in a region and have an impact on COVID-19 propagation. As a respiratory disease, COVID-19 spreads based on human to human interactions. Lau et al. [15] analyzed air traffic data and found a strong linear correlation between domestic COVID-19 cases and passenger volumes for regions within China and a significant correlation between international COVID-19 cases and passenger volumes. Sokadjo and Atchadé [16] found similar results for air traffic using datasets of cases from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University and air transport (passengers carried) from the World Bank. Ayan et al. [17] used cellular network traffic data and a Markovian model that captures the mobility of individuals across municipalities of the city and defined a mobility-aware COVID-19 case prediction model that predicted the number of cases for the following week. The study showed that adding estimates for human mobility as input data was able to improve the accuracy of COVID-19 forecasting models.

This paper presents a new method to compare OD traffic matrix estimations based on the prediction accuracy of a machine learning model trained to estimate one-week ahead COVID-19 incidence values. The spatial granularity of the model will be based on geographic areas. A graph is used to capture the interactions among different areas. A machine learning model able to explore graph information is used to estimate one-week ahead COVID-19 incidence data. COVID-19 data is also divided into the same geographical topology. A one-week ahead horizon is used in order to take the virus incubation time into account. Existing machine learning models to estimate OD traffic matrices such as [6] require extensive ground truth data to train the models which is not always available. Using the mutual information between human mobility and the spread of the COVID-19 virus, the model proposed in this paper uses the already gathered information for COVID-19 monitoring in order to assess the likelihood of different estimations for OD traffic matrices to coincide with real unknown human mobility flows. The proposed model generates daily or weekly graphs based on COVID-19 measured data and OD traffic estimations based on theoretical models and data from traffic sensors. The proposed model defines a novel architecture to extract both the spatial and temporal patterns governing the spread of the COVID-19 virus. The primary objective is to have a method to estimate the accuracy for OD traffic estimations based on traffic sensor information not requiring extensive and expensive training data but using the already COVID-19 gathered data. A secondary objective is to show that the spatial spread of the COVID-19 virus can be better estimated using traffic data and mobility estimations. A limitation of the proposed model is that it is only applicable to estimate OD traffic matrices for regions (such as provinces) for which COVID-19 data is aggregated. This limitation, on the other hand, alleviates the need for having the exact knowledge of the road network and is also present in previous studies such as [6].

III. DEEP GNN MODEL BASED ON O-D TRAFFIC ESTIMATIONS

This section provides the details for the machine learning model proposed to evaluate prior OD traffic flow estimates. The model uses a Graph Neural Network (GNN) to extract mobility aware spatial patterns and a subsequent Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) to extract temporal patterns controlling the spread of the COVID-19 virus. The model will be used to forecast one-week ahead COVID-19 incidence values in the different regions of the graph. The hypothesis is that better OD traffic flow estimates will provide more accurate input data to the model and more accurate input data will make it possible to generate more accurate predictions (section V will capture the data analysis to support that hypothesis). If the model is able to extract the information from the input data it will provide more accurate estimates for the one-week ahead COVID-19 incidence values.

The first two sub-sections are used to present the basic equations modeling GNN and LSTM based models. The proposed GNN-LSTM model is the presented in sub-section C.

A. GRAPH NEURAL NETWORKS (GNN)

Graph Neural Networks (GNNs) directly operate on graph-structured input $G = \{V, E, W\}$ [6]. V represents a set of vertices, E the edges connecting them and W a weighted adjacency matrix. Several types of GNNs have been proposed for processing the information in graphs. Gilmer et al. provide a detailed review over existing graph neural network models [18] in which different models are grouped based on the internal graph operations.

Zhou et al. proposed a two phase GNN model called Message Passing Neural Network (MPNN) [19]. The first phase aggregates the information from adjacent nodes using a message passing function. The second phase performs an update of the nodes' feature vectors based on the previous features and the messages passed. The basic operations describing a MPNN are captured in Equation 1.

$$m_v^{t+1} = \sum\nolimits_{u \in N_v} M_t\left(h_v^t, h_u^t, e_{uv}\right),$$
$$h_v^{t+1} = U_t\left(h_v^t, m_v^{t+1}\right) \tag{1}$$

where h_v^t is the feature vector in node v at time t, e_{uv} represent the features for the directed edge connecting node u with node v, N_v are the neighbor nodes for node v, M_t is the message function to aggregate the messages m_v^t for the computation of the update in the features for node v, and U_t is an update function to compute the next feature vector for each node based on the current feature vector and the messages from the connected nodes.

B. LONG SHORT-TERM MEMORY (LSTM) MODEL

Long Short-Term Memory (LSTM) [20] is a type of Recurrent Neural Network (RNN) machine learning model designed to extract patterns in data sequences such as time series. LSTMs have a state c, and generate an output o_t for each time step based on the input x_t, the output of the previous time step o_{t-1} and the current state c_t information. The model uses three gates (input in_t, output out_t and forget f_t) to control the updates for the state c_t and the output o_t for each time step following Equation 2.

$$\begin{aligned}
f_t &= \sigma_g\left(W_f x_t + W_f' o_{t-1} + b_f\right), \\
in_t &= \sigma_g\left(W_{in} x_t + W_{in}' o_{t-1} + b_{in}\right), \\
out_t &= \sigma_g\left(W_{out} x_t + W_{out}' o_{t-1} + b_{out}\right), \\
\tilde{c}_t &= \sigma_c\left(W_c x_t + W_c' o_{t-1} + b_c\right), \\
c_t &= f_t * c_{t-1} + in_t * \tilde{c}_t, \\
o_t &= out_t * \sigma_h\left(c_t\right),
\end{aligned} \tag{2}$$

where W_α are weight matrices (for $\alpha = f, in, outorc$), b_α represent biases, σ_g, σ_c and σ_h are activation functions (sigmoid for σ_g, hyperbolic tangent for σ_c and hyperbolic tangent or identity for σ_h), and $*$ is a component wise multiplication operation.

C. GNN-LSTM MODEL FOR OD TRAFFIC ASSESSMENT

OD traffic matrices capture the movements from each origin to each destination over a period of time. Depending on the granularity, each origin and destination can represent a road intersection, a road link or an entire area in the map. The information describing the particular features of each origin and destination, together with the traffic flows among them, can be captured in a graph structure. This section uses a graph format to capture both COVID-19 incidence and traffic flow information data for a set of regions. Each region will be both a source and a destination for traffic with other regions and will be represented as a node in the graph. The COVID-19 incidence data for each region will be stored at each node. The edges in the graph will contain estimations for the OD traffic between each pair of regions. The graph will contain directed edges in which flow estimations does not have to be symmetric.

A GNN-LSTM machine learning model that extends the proposal in [21] has been designed in order to predict the COVID-19 incidence one-week ahead (considering the incubation time of the virus) for each node (region). A graph is generated with the COVID-19 incidence information per day. The GNN is used to explore the entire graph and extract the spatial features which are then fed into an LSTM based RNN in order to extract the temporal patterns and generate a prediction for the one-week ahead COVID-19 incidence data. The model is captured in Figure 1.

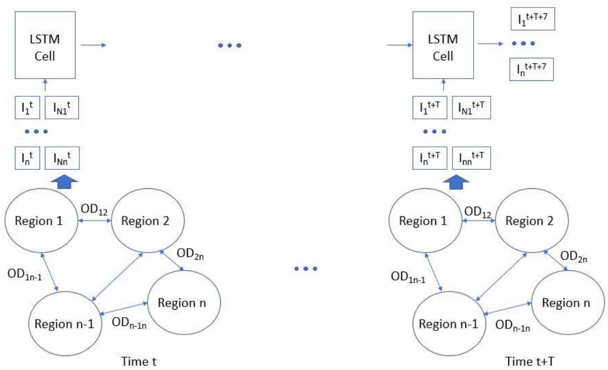

FIGURE 1. GNN-LSTM model for OD traffic assessment.

The GNN in the model in Figure 1 defines following message passing aggregation function for the neighbors as captured by Equation 3.

$$\mathrm{M_t}\left(\mathrm{h_v^t}, \mathrm{h_u^t}, \mathrm{e_{uv}}\right) = \mathrm{k} * \mathrm{e_{uv}} * \mathrm{h_u^t}$$
$$m_v^t = \frac{1}{N_u} \sum_{u \in N_u} \left(M_t\left(h_v^t, h_u^t, e_{uv}\right)\right) \tag{3}$$

where the edge $e_{uv} = OD_{uv}$, with OD_{uv} being an estimation of a prior OD traffic flow between the nodes (regions) u and v [7] and k is a regularization constant that represents the percentage of the traffic flow that contributes to the propagation of the virus (based on the transmissibility of the COVID-19 virus). Each region will spread the virus over connected regions based on the product of the current COVID-19 incidence values and the mobility of the population with other destinations. The aggregated messages for node v from all neighbors N_u are the average of the messages from each neighbor.

The update function in the GNN is defined in Equation 4.

$$U_t\left(h_v^t, m_v^t\right) = concat\left(h_v^t, m_v^t\right) \tag{4}$$

where the average messages from all the neighbors m_v^t are concatenated with the COVID-19 incidence for node v.

The output of the GNN, for each node is then captured in Equation 5.

$$h_v^{t-out} = concat\left(h_v^t, m_v^t\right) = concat\left(I_v^t, I_{Nv}^t\right) \tag{5}$$

where I_v^t is the COVID-19 incidence at node v for time t and I_{Nv}^t is the averaged incidence of the neighbors at time t weighted by the estimated OD traffic flows.

The outputs for the GNN neural network will then be fed into a LSTM model with n_c memory units which will be trained to estimate the one-week ahead COVID-19 incidence values. A simplified similar model has been successfully used in order to estimate traffic flows in road links based on the information of traffic on other links weighted by the distance between links in [21]. We propose a model that extends the graph information with COVID-19 data and uses prior estimates for OD traffic flows as weights in the aggregation function in Equation 3. Since the COVID-19 virus is propagated based on human to human interactions, good prior estimates for the OD traffic matrix will help the model in Figure 1 to achieve better predictions for one-week ahead COVID-19 incidence values. A similar model combining a GNN connected to an LSTM for electroencephalography (EEG) signal classification has been proposed in [22]. The model used electrodes as nodes in the graph and an adjacency matrix based on the k nearest neighbors for each electrode showing that the model offered promising results for emotion recognition, and motor imagery decoding. Other variants for the GNN model such as the one proposed in [23] could be used as a future work. A similar GNN+LSTM model has also been recently used in [24] in order to anticipate COVID-19 incidence data for Brazil, also showing promising results.

IV. DATASETS

The model in Figure 1 has been applied to data in Spain for the entire 2021, in which data on new COVID-19 infections for each region (province) was reported every day. Two major datasets have been used to validate the model and apply it to different prior OD matrix estimates: the COVID-19 incidence data from the Carlos III Health Institute [25] and the traffic flow data on Spanish roads from the Transport Ministry [26].

The COVID-19 incidence data values in [25] are obtained from the declaration of COVID-19 cases to the National Epidemiological Surveillance Network (RENAVE) through the computer platform via Web SiViES (Surveillance System of Spain) managed by the National Epidemiology Center (CNE). The dataset contains daily reported cases for different age ranges and gender, per province. In order to simplify the model, the data has been aggregated per day and province. Each node v in the graph will contain a single state variable with the aggregated reported cases for that node (using desegregated data will be studied as a future work). Only new reported COVID-19 infections are considered. The dataset also contains information for hospitalizations and deaths which could be used to enhance the proposed model as a future work.

The traffic data [26] captures the monthly evolution of the traffic registered in the traffic meters in 635 locations of the Spanish road network in 44 provinces. The dataset also contains the estimated total traffic at the end of each year per traffic meter. The traffic data from 2021 will be used to generate OD prior matrices to train and validate the model. Since the temporal granularity of the traffic data is different from the COVID-19 incidence data, OD prior estimates will be replicated for all the input graphs for that period.

V. INITIAL DATA ANALYSIS

In this section we analyze the hypothesis that mobility and COVID-19 incidence data are correlated using datasets containing information for Spain in 2021. The machine learning models proposed in this paper will use the mutual information among these two variables in order to infer better mobility estimations based on the ability to learn patterns leading to better estimates of COVID-19 incidence data.

In this section, we also analyze the limitations of using road traffic sensors as proxy variables to estimate human mobility.

A. MOBILITY DATA AND COVID-19

COVID-19, as a respiratory virus, spreads based on human to human interactions (co-located people over a sufficient amount of time). Human mobility allows infected people to disseminate the virus. The impact of mobility in the spread of the virus has been captured in several studies. Alessandretti [27] provided a review study on the causal mechanisms that lead to link human travel with close human interactions capable of spreading the COVID-19 virus. The authors found causal patterns at the beginning of the COVID-19 pandemic. Similar results were presented in [28] where origin–destination travel demand and aggregate mobility inflow at each US county from March 1 to June 9, 2020 were computed, and a positive relationship between mobility inflow and the number of infections during the COVID-19 onset was found. The authors in [29] found a relationship between human mobility and the virus spread in previous studies although the relationship presented temporal and spatial heterogeneity. For the particular case of Spain, the study in [30] confirmed that daily new Coronavirus COVID-19 cases were directly related to mobility habits 14 days before for the first months of the pandemic. In this subsection we study the correlation of mobility and COVID-19 propagation for the data in 2021 which is the one used in this paper.

In order to assess the impact of mobility on the spread of the COVID-19 virus, the dataset in [31] has been used. This dataset analyses mobility data obtained from cellular phone aggregated flows for the different regions in Spain in 2021. The mobility data will be analyzed together with the COVID-19 incidence data for each region in Spain [25]. Data in [25] shows two major COVID-19 waves of infections in Spain in 2021. The first wave was influenced by the Christmas period involving mobility patterns that are not captured in [31]. The second wave of infections has therefore been used in this section. The Pearson correlation index for the 5 weeks before and after the peak of infections for the second wave for mobility and COVID-19 incidence time series has been calculated in order to analyze the mutual information both in the case of the increase and decrease of COVID-19 infections. The results are presented in Table 1 showing a positive correlation for mobility patterns and COVID-19 new infections.

TABLE 1. Pearson correlation coefficient.

	Before peak of infections	After peak of infections
Pearson correlation	0.83	0.87

Instead of using data from cellular phones, we propose a different approach based on the information obtained from traffic sensors in this paper making use of the mutual information for the mobility and COVID-19 incidence data in Spain in 2021.

B. TRAFFIC ANALYSIS BY DIFFERENT MEANS OF TRANSPORTATION

The mobility flows in this paper are obtained from traffic sensors as a proxy variable to estimate total mobility in Spain. The data in [32] for 2021 shows that around 90% of the total mobility for passengers (human mobility) is road traffic based. Table 2 captures the results.

TABLE 2. Internal mobility in spain by means of transport in 2021 (Billions).

	Road	Train	Plain
Passengers-km	354	17.4	26.1

Together with passenger mobility, traffic sensors also capture freight traffic. The dataset in [26] shows that freight traffic was below 10% of passenger traffic. In this paper, we are going to use the outputs for each traffic sensor as a proxy variable for human mobility considering the majority of human mobility in Spain in 2021 used road transportation and that freight traffic is an order of magnitude smaller than passenger traffic. A more detailed analysis by means of transport will be done in a future study.

VI. MODELS FOR ESTIMATING PRIOR OD TRAFFIC MATRICES

Different methods have been used to generate prior OD matrices such as [8] and [9].

Willumsen [8] defined a gravity model for prior OD estimations based in Equation 6.

$$T_{ij} = b_i * O_i * D_j * c_{ij}^{-d}, \tag{6}$$

where T_{ij} is the estimate for the traffic from origin i to destination j, O_i represents the traffic from origin i, D_j represents the traffic to destination j, c_{ij} represents the cost for travelling from i to j, c_{ij}^{-d} represents the traffic attraction and b and d are constants.

Tsekeris and Stathopoulos [9] presented a dynamic gravity model in which OD traffic data is optimized based on the entropy maximization criterion following Equation 7.

$$T_{ij} = A_i * \tilde{O}_i * B_j * \tilde{D}_j * e^{-\gamma c_{ij}}, \quad (7)$$

where T_{ij} is the OD traffic computed based on the estimation or total traffic from origin $\tilde{O}_i$ and destination $\tilde{D}_j$ and a travel cost function c_{ij} which leads to a traffic attraction exponential function $e^{-\gamma c_{ij}}$. A_i and B_j are balancing parameters.

In this section we describe 3 prior OD matrix estimation models based on previous studies and a baseline model which will be used to compare the OD estimation models based on the results from the GNN-LSTM in Figure 1. The information from traffic meters described in the previous section will be used. Each traffic meter is geo-located in a particular location inside a particular region (province). Each province is a node in the graph and all the regions generate and attract traffic from connected provinces.

A. GRAVITY MODEL WITH A TRAFFIC ATTRACTION FUNCTION BASED ON THE INVERSE OF THE DISTANCE

Each node (province) in the graph may contain several traffic meters. A modified version of Equation 6 is used, adapted to regions comprising several traffic meters. Origin and Destination traffic is computed adding the traffic measured by all the traffic sensors in the region according to Equation 8 (other approximations including the population of each region will be studied in future work).

$$O_i = \sum\nolimits_{s\in N_i} T_s, D_j = \sum\nolimits_{s\in N_j} T_s, \quad (8)$$

where N_i represents the number of traffic meters in the origin region and N_j the number of meters in the destination.

The cost in Equation 6 is used to estimate the attraction for the traffic between two regions as described in Equation 9.

$$\text{attraction}_{ij} = c_{ij}^{-d} = \text{dist}_{CiCj}^{-1} \quad (9)$$

where *Ci* and *Cj* represent the "center" of the origin and destination regions and *dist* is the Euclidean distance. Since the majority of the traffic is generated near the capital city of each region, we have set the traffic "center" for this first model in the geographic location of the capital.

B. GRAVITY MODEL WITH EXPONENTIAL BASED TRAFFIC ATTRACTION FUNCTION

The second model analyzed in the current study will use a similar gravity model as the one in the previous section but using an exponential attraction function as captured in Equation 7. The cost for traffic going from region i to region j is captured in Equation 10.

$$\text{attraction}_{ij} = e^{-\gamma \text{dist}_{CiCj}} \quad (10)$$

where *Ci* and *Cj* represent the "center" of the origin and destination regions and *dist* is the Euclidean distance. The parameter γ is set to

$$\frac{1}{min\left(\text{dist}_{CiCj}\right)}.$$

The traffic "center" for this second model is also located in the geographic location of the capital city (assuming that the majority of the traffic for a particular region is based in the capital city and its surroundings).

C. DIRECTED CENTER OF TRAFFIC MASS MODEL FOR TRAFFIC FLOWS AMONG DIFFERENT REGIONS

The third model proposed for OD prior matrix estimation for region to region (province to province) traffic flows will estimate the amount of traffic by computing the center of directed "traffic mass" according to Equation 11.

$$T_{ij} = \frac{1}{\left\|\overrightarrow{C\imath C\jmath}\right\|} \sum\nolimits_{s\in N_i} \frac{T_s \overrightarrow{C\imath Cs} \cdot \overrightarrow{C\imath C\jmath}}{\left\|\overrightarrow{C\imath C\imath}\right\|}, if\, \overrightarrow{C\imath Cs} \cdot \overrightarrow{C\imath C\jmath} > 0 \quad (11)$$

where T_s captures the traffic counted by sensor s, $\overrightarrow{C\imath Cs}$ represents a vector connecting the "center" of the origin province i and the traffic sensor s and $\left\|\overrightarrow{C\imath Cs}\right\|$ its norm, N_i is the number of traffic sensors in province i, $\overrightarrow{C\imath C\jmath}$ is the vector connecting the "centers" for provinces i and j, $\left\|\overrightarrow{C\imath C\jmath}\right\|$ its norm and $\cdot$ represents the dot product. The capital cities are also selected for the "center" of each province. Equation 10 only considers traffic sensors in region i that fulfill that $\overrightarrow{C\imath Cs}\cdot\overrightarrow{C\imath C\jmath} > 0$, considering that sensors with $\overrightarrow{C\imath Cs}\cdot\overrightarrow{C\imath C\jmath} < 0$ are more likely to capture traffic exchanged with other regions. The dot product $\overrightarrow{C\imath Cs}\cdot\overrightarrow{C\imath C\jmath}$ is used to project the amount of T_s in the flow going in the direction form province i to province j.

The OD prior estimation following Equation 10 assigns a higher probability to traffic from sensors in region (province) i to flow to region (province) j to those sensors which are closer to the "center" (considering the capital city as the traffic center) of region (province) j.

D. BASELINE MODEL

A baseline (traffic agnostic) method is also defined in order to assess the model gain when using the previous OD traffic matrix estimation methods in order to achieve a better COVID-19 forecasting accuracy.

The baseline model will assume that the traffic flows from each region will be equally shared with all neighbor regions according to Equation 12.

$$\begin{aligned} T_{ij} &= 1 \text{ for neighbour regions} \\ T_{ij} &= 0 \text{ for non} - \text{neighbou regions} \end{aligned} \quad (12)$$

All connected regions (provinces) will share the same amount of normalized traffic volumes with each other and no traffic with non-connected regions without taking the traffic sensor data into account.

VII. RESULTS

This section captures the results when using the GNN-LSTM machine learning model in Figure 1 to forecast one-week ahead COVID-19 incidence values for each province in Spain during the entire 2021 using the datasets described in section IV. The 4 methods captured in section VI will be used in order compute the values for the edges of the graph (e_{uv} in Equation 3) as estimates of the Origin-Destination (OD) flows among different provinces. Better OD flow estimations will provide more accurate input information to the GNN-LSTM machine learning model so that the output of the model is able to generate more accurate predictions.

A. NORMALIZED ADJACENCY MATRICES

In order to speed up the training of the machine learning model in Figure 1, the values for the OD estimates are normalized following Equation 13 when feeding the input values to the GNN-LSTM model in Figure 1.

$$e_{ij} = \frac{A_{ij}T_{ij}}{max\ (A_{ab}T_{ab})\ \forall a, b} \tag{13}$$

where A is an adjacency matrix, A_{ij} each element in the adjacency matrix and e_{ij} captures the estimated OD traffic between regions i and j. $A_{ij} = 1$ for connected regions i and j and $A_{ij} = 0$ for the non-connected ones. Since the traffic attraction values used to compute prior OD matrices in Equations 7 and 9 decrease with the distance, in order to simplify the graph for the model in Figure 1, only the nearest (in distance) provinces are considered to be connected with $A_{ij} = 1$ (other adjacency matrices will be studied in future work).

The road sensors in the traffic dataset [26] are deployed in the Spanish roads are highlighted in Figure 2. The traffic sensors in Figure 2 are shown inside the different Spanish provinces (which will be the nodes in the graph). The provinces with $A_{ij} = 1$ for the province of Segovia (as an example) are captured in Figure 3 (a threshold of 150 km has been used). A greyscale visualization of the traffic attraction to the province of Segovia (represented in red) following Equation 9 is captured in Figure 4. Provinces with $A_{ij} = 0$ show a low attraction value (darker values).

The estimated OD traffic flows T_{ij} are normalized in Equation 12 using a min-max scaler function so that the edge values e_{ij} have a maximum value of 1 to improve the convergence of the training of the GNN-LSTM machine learning model.

B. LEAVE ONE-WAVE-OUT CROSS-VALIDATION

The spatio-temporal spread of the COVID-19 has generated waves of infections. Figure 5 shows the aggregated (all provinces, all ages, all genders) reported cases of new COVID-19 infections in Spain during 2021. Three major waves are shown: one at the beginning of the year, a second one around August and a final one by the end of the year. The latest wave was driven by the Omicron COVID-19 variant which has a higher infectivity, generating a higher number of COVID-19 reported cases.

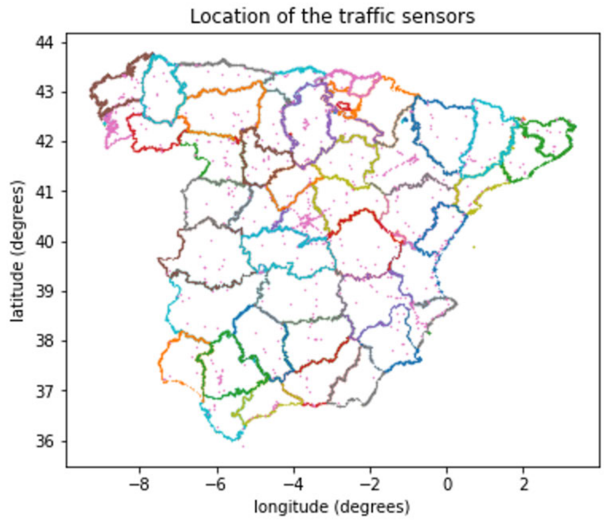

FIGURE 2. Spanish provinces and the location of the traffic sensors.

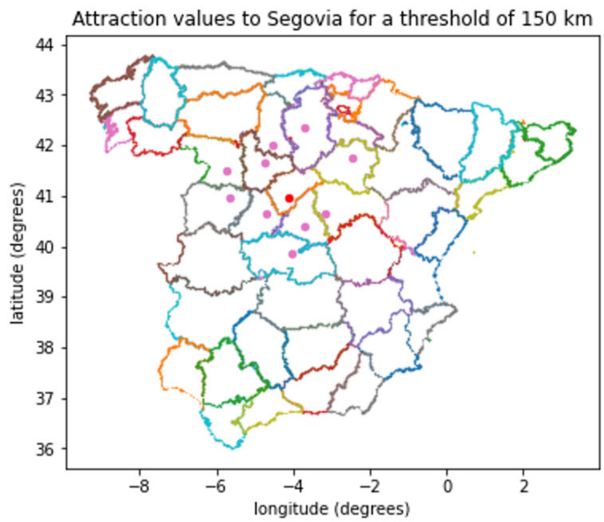

FIGURE 3. Adjacency example (province of Segovia).

In order to validate the accuracy of the model in Figure 1, a leave one wave out approach is followed. The data for the first wave is separated for testing and the rest of the data for training. The origin of the first wave is the first of January, 2021, and the end has been estimated as February 20, 2021. The information from February 21 to December 31, 2021 is used for training. The model will learn the propagation patterns of the COVID-19 virus from the second and third waves and use the learned patterns to estimate the spread of the first wave.

For each configuration of the model in Figure 1 (with different number of memory units in the LSTM cells) and for each OD matrix estimation method, a total of 10 different training optimizations (executions) are performed, using random shuffling for the data samples in order to train the model. The average values for the accuracy of the model are

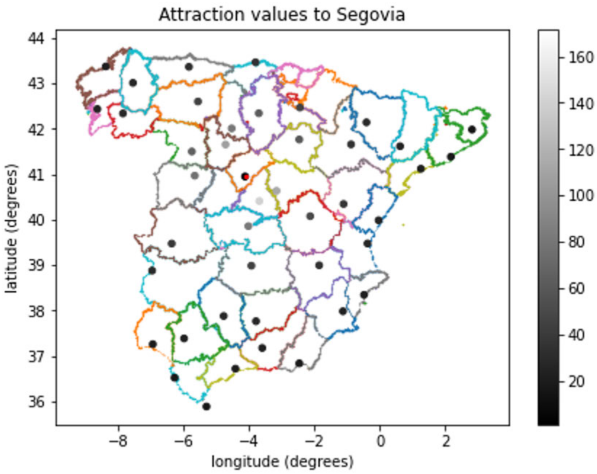

FIGURE 4. Attractions values for traffic to the province of Segovia.

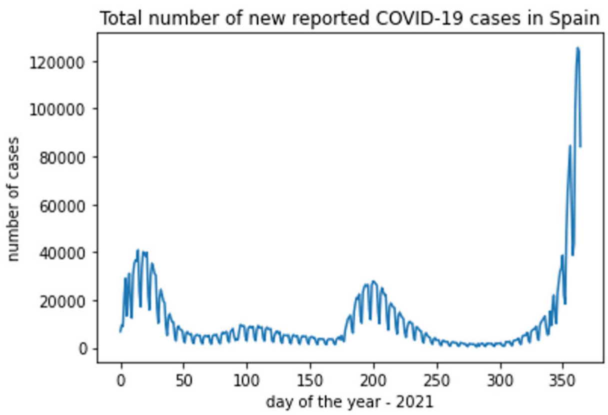

FIGURE 5. New daily aggregated COVID-19 cases in Spain (2021).

then calculated. This is done to minimize the effect of the numeral optimization process has on the output of the model.

Two error metrics commonly found in related studies will be used to validate the accuracy of the model: the Mean Absolute Error (MAE) and the Mean Squared Error (MSE). Taking the incubation period for the COVID-19 virus into account, a one-week ahead estimation will be used (the model will try to predict one-week ahead COVID-19 incidence data). The model will use the information of the previous 3 weeks to minimize the one-week ahead estimation error.

C. GNN-LSTM MODEL OPTIMIZATION AND ABLATION STUDY

The number of memory units in the LSTM layer in the GNN-LSTM model in Figure 1 is a parameter that can be used to control the complexity of the model. High complexity model configurations are able to learn more intricate patterns in the data reducing the bias in the predictions. Low complexity models are able to learn basic patterns but tend to generate a higher bias for datasets in which patters are complex. High complexity models are more likely to generate higher variances (overfitting) in predicted values. The leave one out validation approach presented in the previous section will be used in order to validate that the trained model according to different numbers of LSTM memory units generalize to new data (not incurring in overfitting issues).

Figure 6 shows the MAE errors for 5 different values for the number of memory units in the LSTM layer for the 4 approaches presented in this paper in order to provide OD prior estimates. Figure 7 captures the values for the MSE errors for the same configurations for the number of memory units. In both cases, the baseline model (assuming that all connected provinces exchange the same amount of normalized traffic) produces an optimal performance for 16 memory units in the LSTM layer, meaning that low complex configurations for the machine learning model are not able to learn from real traffic data in order to estimate the spread of the virus among different provinces. For more complex configurations of the model (LSTM memory units equal or greater than 32) the model in Figure 1 is able to extract the information from the traffic data and optimize the predictions for one-week ahead COVID-19 incidence data. Figure 6 shows that the Mean Absolute Error (MAE) for the directed center of traffic mass achieves the optimal performance while the other two methods (the gravity models based on traffic attraction inversely proportional to distances or using exponential attraction values) show a similar performance. The directed center of traffic mass method is able to use the available traffic information in greater detail. The method incorporates the traffic volumes measured considering the exact location of the traffic sensor and the relative location within the two closest provinces. It also incorporates a likelihood estimation for the traffic to cross the border of the two provinces and filters the traffic that is unlikely to do it. The Mean Square Errors (MSE) show similar results except for the case of 32 LSTM memory units and the gravity model based on exponential costs which achieves the optimal accuracy.

The results in Figure 6 and Figure 7 show that using the information of the geographical location of traffic meter sensors in an OD traffic flow estimation method is able to outperform simpler aggregated methods that add the traffic in each province. The directed center of traffic mass proposed in this paper assumes a probability for the traffic measured at a particular traffic meter to cross the border with a neighbor province based on the distance of the projected sensor location to the neighbor province. Other probability methods would be studied as a future work in order to assess other OD prior matrix computations.

In order to assess the importance of both parts of the model in Figure 1 (the GNN and the LSTM parts of the model) in the overall accuracy, an ablation study has been carried out. The results are shown in Table 3. Both MSE and MAE values are provided for average executions for all the OD estimation models and the number of memory units in the LSTM layer as used in Figures 6 and 7. Table 3 shows that removing the GNN part of the model has a higher impact on the accuracy of the model (showing the importance of the information provided by the traffic flow estimations on the overall performance of the model).

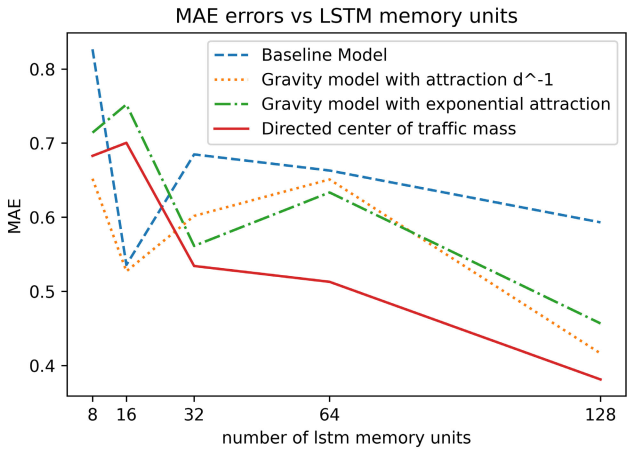

FIGURE 6. Average (10 executions) MAE errors for the validation set (first wave) vs LSTM memory units.

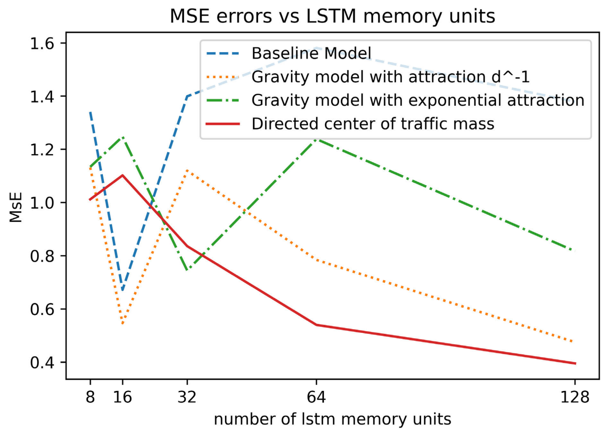

FIGURE 7. Average (10 executions) MSE errors for the validation set (first wave) vs LSTM memory units.

D. MODEL GAIN BY THE DIFFERENT OD TRAFFIC ESTIMATIONS

In order to assess the model gain (the increase in the accuracy when using the different OD traffic estimation methods compared with the baseline model), the accuracy of the model in Figure 1 has been averaged for low complexity configurations (up to 16 LSTM memory units) and high complexity configurations (from 32 to 128 LSTM memory units). The averaged MAE results are shown in Figure 8. The averaged MSE values are captured in Figure 9.

For low complex configurations for the machine learning model in Figure 1, using traffic data from traffic meters as edge information in the input graph does not provide a clear gain for the model accuracy (the directed center of traffic mass estimator shows similar accuracy values as the baseline model, the gravity model based on inverse distance attraction values shows a small gain, while the model based on exponential cost throws model losses instead of gains). For higher model complexity configurations, using traffic sensor data and a method to estimate prior OD flows, allows the model in Figure 1 to optimize its performance. The directed center of traffic mass proposed method is able to outperform aggregated traffic methods for both MAE and MSE error values.

TABLE 3. Average Accuracy of the Predictions when removing one part of the model.

	Complete model	Eliminating the GNN	Eliminating the LSTM
MAE	0.624	0.871	0.732
MSE	1.036	1.551	1.083

Table 4 captures the model gains for the different prior OD traffic flow estimations when using both MAE and MSE as the accuracy values for the model predictions (COVID-19 incidence values one week ahead). The model gain is defined as the prediction error achieved the baseline model divided by the error of the compared model (smaller prediction errors will lead to gains>1). The model gains for the MSE are higher since the majority of the prediction errors are in the range $[-1\ 1]$ for which the squares of the values are smaller than the absolute values (so that the average of the errors tends to be smaller for the MSE error and the gains are therefore higher).

E. COMPARING RESULTS WITH PREVIOUS STUDIES

The major contribution of this paper is to propose a model, using the mutual information between human mobility and COVID-19 new infections, to compare different methods to compute prior OD matrices from traffic sensor data. The geographical information of the area under study is represented as a graph. The COVID-19 incidence for each node in the graph is added as an attribute to the node. Each theoretical method to compute prior OD matrices from traffic sensor data will be used to generate mobility estimates for the edges of the graph. A GNN is then used to extract interaction patterns from the graph. A time sequence of processed graphs after each GNN is then applied to a LSTM model to try to estimate one-week ahead COVID-19 incidence for each node (region). Theoretical models able to better predict OD matrices will be able to provide graphs which are closer to real data and therefore to better extract COVID-19 and human mobility mutual information. Figures 8 and 9 capture the accuracy results when using 3 different theoretical methods to estimate OD matrices as compared to a baseline model. In this section, we go a step further and validate that adding mobility data, the accuracy in estimating one-week ahead COVID-19 data improves. The approach used is to compare results with previous state of the art models to estimate short-term COVID-19 new cases that do not use mobility data.

Different machine learning models have been used in order to provide short term estimations for COVID-19 incidence values. An extensive comparison of models proposed in previous research studies is captured in [33]. The authors in [33] proposed a spatio-temporal model to estimate one-week ahead COVID-19 new cases and compared the results with previous studies for similar estimations. The

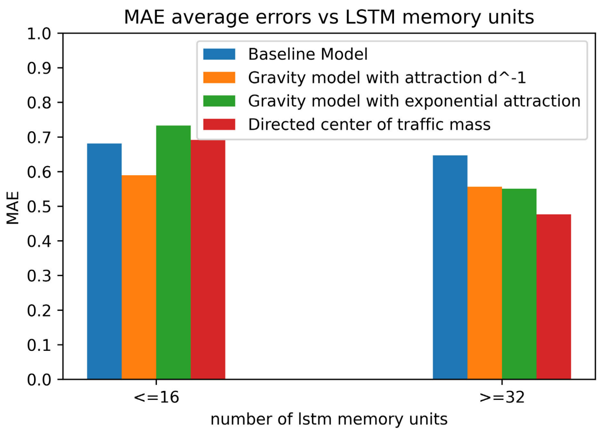

FIGURE 8. **MAE errors for low and high complexity model.**

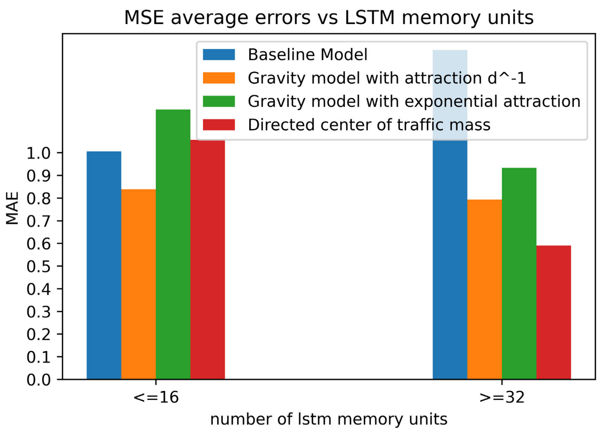

FIGURE 9. **MSE errors for low and high complexity model.**

TABLE 4. **Model gain for the different od estimation methods.**

	BASELINE	D^1	EXPONENTIAL	DCTM
MAE	1	1.203	1.192	1.385
MSE	1	2.062	1.628	2.692

machine learning models did not consider human mobility estimations in order to improve results. The best performing model for one of the Spanish provinces (Community of Madrid) combined a CNN to process the spatial distribution on new COVID-19 cases with a subsequent LSTM for time series estimations. The model achieved an average MSE value of 3.7 for the different areas inside the Community of Madrid for the data for 2021. The models proposed in this paper use a GNN instead of a CNN which is able to both adapt to the irregular geometry of each province and at the same time to add the mobility information in the edges of the graph. The optimal MSE for the best performing model in this paper is 0.6. Adding mobility information is able to provide better estimations as compared to previous models not using traffic estimations (aligned with the results in section V, traffic data provides information about COVID-19 spreading).

VIII. CONCLUSION

This paper has proposed a Graph Neural Network (GNN) to analyze the traffic exchange estimations among provinces in Spain. OD traffic flow estimations are used to provide values for the edges of the graph. The output of the GNN is fed into a Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) in order to analyze temporal patterns of the output of the processed graph. COVID-19 incidence values for 44 regions providing new daily reported cases during 2021 and traffic counts from 635 traffic meters in different road segments have been used to generate daily graphs (using COVID-19 incidence data as node information and OD traffic estimations as edge data).

The major contribution of the paper is providing and evaluating a new methodology for estimating the accuracy provided by different OD traffic flow estimation methods based on the use of a GNN-LSTM machine learning model and the process controlling the spread of the COVID-19 virus. Since COVID-19 virus is mainly propagated from people to people interactions, better OD traffic flow estimates will help the machine learning model to better incorporate the COVID-19 incidence data from neighbor regions into one-week ahead COVID-19 incidence estimations. Traffic and COVID-19 mutual information for the datasets used in this paper for Spain in 2021 is also analyzed in order to validate the hypothesis that better OD estimations will allow the ML model to better predict COVID-19 upcoming values.

The results for 2 distance base gravity methods to estimate prior OD traffic flows and a new method based on the computation of the directed center of traffic mass show that the GNN-LSTM machine learning model improves the results of a traffic agnostic baseline model when the complexity of the model is sufficient to take the input information in the edges of the graph into account. The directed center of traffic mass method uses a higher detail in the traffic information, incorporating the geographical location of each traffic sensor in the OD flow estimations based on a probability estimate that assigns higher probability values to traffic closer to the border between two regions (which are more likely to cross the border). The directed center of traffic mass method achieves the optimal model gains for both the MAE and MSE errors.

The major application of the results of this paper is a novel mechanism to validate prior OD traffic matrices.

As future work, other OD traffic flow estimation methods will be used in order to assess the model gains for different model configurations.

One of the limitations of the research in this paper is that the proposed model has mainly been used as a validation tool for different OD traffic flow estimation methods. As a future work, we will apply the model for improving the estimated traffic flows based on the optimization of the accuracy of the model, using the values of the edges of the graph as trainable variables in the optimization process.

REFERENCES

[1] S. Afandizadeh Zargari, A. Memarnejad, and H. Mirzahossein, "Hourly origin–destination matrix estimation using intelligent transportation systems data and deep learning," *Sensors*, vol. 21, no. 21, p. 7080, Oct. 2021.

[2] P. Wang, J. Lai, Z. Huang, Q. Tan, and T. Lin, "Estimating traffic flow in large road networks based on multi-source traffic data," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 9, pp. 5672–5683, Sep. 2021.

[3] H.-J. Cho, Y.-J. Jou, and C.-L. Lan, "Time dependent origin-destination estimation from traffic count without prior information," *Netw. Spatial Econ.*, vol. 9, no. 2, pp. 145–170, Jun. 2009.

[4] Z. Liu, Y. Liu, Q. Meng, and Q. Cheng, "A tailored machine learning approach for urban transport network flow estimation," *Transp. Res. Part C, Emerg. Technol.*, vol. 108, pp. 130–150, Nov. 2019.

[5] T. Pamuła and R. Sochowska, "Estimation and prediction of the OD matrix in uncongested urban road network based on traffic flows using deep learning," *Eng. Appl. Artif. Intell.*, vol. 117, Jan. 2023, Art. no. 105550.

[6] Z. Liu, Z. Liu, and X. Fu, "Dynamic origin-destination flow prediction using spatial–temporal graph convolution network with mobile phone data," *IEEE Intell. Transp. Syst. Mag.*, vol. 14, no. 5, pp. 147–161, Sep. 2022.

[7] P. Timms, "A philosophical context for methods to estimate origin? Destination trip matrices using link counts," *Transp. Rev.*, vol. 21, no. 3, pp. 269–301, Jul. 2001.

[8] L. G. Willumsen, "Simplified transport models based on traffic counts," *Transportation*, vol. 10, no. 3, pp. 257–278, Sep. 1981.

[9] T. Tsekeris and A. Stathopoulos, "Gravity models for dynamic transport planning: Development and implementation in urban networks," *J. Transp. Geography*, vol. 14, no. 2, pp. 152–160, Mar. 2006.

[10] S. Sánchez-Cambronero, F. Álvarez-Bazo, A. Rivas, and I. Gallego, "Dynamic route flow estimation in road networks using data from automatic number of plate recognition sensors," *Sustainability*, vol. 13, no. 8, p. 4430, Apr. 2021.

[11] X. Yang, Y. Lu, and W. Hao, "Origin-destination estimation using probe vehicle trajectory and link counts," *J. Adv. Transp.*, vol. 2017, pp. 1–18, Jun. 2017.

[12] Z. Wu, M. Huang, A. Zhao, and Z. Lan, "Urban traffic planning and traffic flow prediction based on ulchis gravity model and Dijkstra algorithm," *J. Phys., Conf. Ser.*, vol. 1972, no. 1, Jul. 2021, Art. no. 012080.

[13] W. Hualan, X. Jiaying, and W. Zeyu, "An estimation method of OD matrix in non-congestion ring expressway," *Transp. Res. Proc.*, vol. 48, pp. 883–889, Jul. 2020.

[14] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 3634–3640.

[15] H. Lau, V. Khosrawipour, P. Kocbach, A. Mikolajczyk, H. Ichii, M. Zacharski, J. Bania, and T. Khosrawipour, "The association between international and domestic air traffic and the coronavirus (COVID-19) outbreak," *J. Microbiology, Immunology Infection*, vol. 53, no. 3, pp. 467–472, Jun. 2020.

[16] Y. M. Sokadjo and M. N. Atchadé, "The influence of passenger air traffic on the spread of COVID-19 in the world," *Transp. Res. Interdiscipl. Perspect.*, vol. 8, Nov. 2020, Art. no. 100213.

[17] N. Ayan, S. Chaskar, A. Seetharam, A. Ramesh, and A. A. D. A. Rocha, "Poster: COVID-19 case prediction using cellular network traffic," in *Proc. IFIP Netw. Conf. (IFIP Networking)*, Jun. 2021, pp. 1–3.

[18] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2017, pp. 1263–1272.

[19] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, Sep. 2020.

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

[21] A. Khodadadi. *Traffic Forecasting Using Graph Neural Networks and LSTM*. Accessed: Mar. 28, 2024. [Online]. Available: https://keras.io/examples/timeseries/timeseries_traffic_forecasting/

[22] P. Gong, P. Wang, Y. Zhou, and D. Zhang, "A spiking neural network with adaptive graph convolution and LSTM for EEG-based brain–computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 1440–1450, 2023.

[23] P. Gong and L. Ai, "Neighborhood adaptive graph convolutional network for node classification," *IEEE Access*, vol. 7, pp. 170578–170588, 2019.

[24] F. H. O. Duarte, G. J. P. Moreira, E. J. S. Luz, L. B. L. Santos, and V. L. S. Freitas, "Time series forecasting of COVID-19 cases in Brazil with GNN and mobility networks," in *Proc. Brazilian Conf. Intell. Syst.*, 2023, pp. 361–375.

[25] *ISCIII, Instituto De Salud Carlos III. COVID-19 Open Data*. Accessed: Mar. 17, 2024. [Online]. Available: https://cnecovid.isciii.es/covid19/

[26] *Ministerio De Transportes, Movilidad Y Agenda Urbana. Monthly Data Traffic Open Data*. Accessed: Mar. 17, 2024. [Online]. Available: https://www.mitma.es/carreteras/trafico-velocidades-y-accidentes-mapa-estimacion-y-evolucion/datos-mensuales-de-trafico/datos-mensuales-de-trafico-en-la-rce

[27] L. Alessandretti, "What human mobility data tell us about COVID-19 spread," *Nature Rev. Phys.*, vol. 4, no. 1, pp. 12–13, Dec. 2021.

[28] C. Xiong, S. Hu, M. Yang, W. Luo, and L. Zhang, "Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 44, pp. 27087–27089, Nov. 2020.

[29] M. Zhang, S. Wang, T. Hu, X. Fu, X. Wang, Y. Hu, B. Halloran, Z. Li, Y. Cui, H. Liu, Z. Liu, and S. Bao, "Human mobility and COVID-19 transmission: A systematic review and future directions," *Ann. GIS*, vol. 28, no. 4, pp. 501–514, Oct. 2022.

[30] J. M. Amoedo, Y. Atrio-Lema, M. D. C. Sánchez-Carreira, and I. Neira, "The heterogeneous regional effect of mobility on coronavirus spread," *Eur. Phys. J. Special Topics*, vol. 231, nos. 18–20, pp. 3391–3402, Dec. 2022.

[31] (2021). *Evolucian De La Movilidad Porombito Geografico Durante El Estado De Alarma Por COVID-19*. [Online]. Available: https://www.ine.es/experimental/movilidad/experimental*4.htm*

[32] Observartorio Del Transporte Y La Logestica En Espaea*. Accessed: Mar. 28, 2024. [Online]. Available: https://observatoriotransporte.mitma.gob.es/movilidad*

[33] M. Muñoz-Organero and P. Queipo-Álvarez, "Deep spatiotemporal model for COVID-19 forecasting," Sensors*, vol. 22, no. 9, p. 3519, May 2022.*

MARIO MUÑOZ-ORGANERO (Senior Member, IEEE) received the M.Sc. degree in telecommunications engineering from the Polytechnic University of Catalonia, Barcelona, Spain, in 1996, and the Ph.D. degree in telecommunications engineering from the Universidad Carlos III de Madrid, Madrid, Spain, in 2004.

He has more than four years of experience working for the telecommunications industry in companies, such as Telefonica Research and Development and Lucent Technologies, Madrid. During a sabbatical leave for the academic course, from 2015 to 2016, he was a Visiting Professor with the University of Sheffield and Nottingham Trent University, U.K. He is currently a Professor of telematic engineering with the Universidad Carlos III de Madrid, where he is also the Head of the Departmental Area. He has authored or co-authored more than 180 research publications (58 indexed in JCR journals) and participated in more than 20 European-funded and Spanish national-funded research projects. His research interests include applied machine learning for sensor data, including traffic sensors in applications for health and mobility.

VÍCTOR CORCOBA-MAGAÑA received the M.Sc. degree in computer engineering from the University of Granada, Barcelona, Spain, in 2010, and the Ph.D. degree in telematics engineering from the Universidad Carlos III de Madrid, Madrid, Spain, in 2014.

He is currently an Associate Professor with the Computer Science Department, University of Oviedo, Spain. He has authored or co-authored more than 80 research publications (23 indexed in JCR journals), an international patent, and participated in more than 14 research projects funded by international, national, and regional/local institutions through competitive calls. He is a member of the Editorial Board of the *Sustainable Cities and Society*. He is also involved in the IoT solutions in the field of health and mobility.