Is Adding More Modalities Better in a Multimodal Spatio-temporal Prediction Scenario? A Case Study on Japan Air Quality

Yutaro Mishima[1,2(✉)], Guillaume Habault[1], and Shinya Wada[1]

[1] KDDI Research, Inc., 2-1-15, Ohara, Fujimino, Saitama Prefecture, Japan
yu-mishima@kddi-research.jp
[2] KDDI Corp., 3-10-10, Iidabashi, Chiyoda-ku, Tokyo, Japan

Abstract. Nowadays, several spatio-temporal datasets are made available for research purposes (e.g., location, traffic or meteorology dataset). These datasets are more and more utilized as multimodal inputs of neural networks in order to perform spatio-temporal predictions. However, there are few methods that include functions, which explicitly capture cross-modal relationships. This lack of information will be a serious problem when more complex modalities and dependencies among modalities will need to be taken into consideration. Considering that in the future more spatio-temporal datasets will be made available, it is of crucial importance to tackle this problem. In this paper, we conduct some preliminary experiments to confirm whether an existing multimodal spatio-temporal network performs better when another modality is added. These experiments compare air quality forecasting performance using a trimodal spatio-temporal dataset. This comparison is realized with several methods and especially one that has been modified to handle multiple modalities. Based on the obtained results, we confirm that prediction performance does not improve when another modality is simply added. Therefore, some methods are required to capture complex cross-modal relationships.

Keywords: Multimodal · Spatio-temporal · Air quality · Location · GPS

1 Introduction

With recent worldwide rise in economic activities, air pollution has grown to be a big issue. Forecasting air quality is decisive for preventing people from health damage caused by air pollution. Therefore, it is important to analyze the relation between air quality and human activities. Such outputs will help regulate air pollution. Meanwhile, people, organizations and governments have made recent efforts in order to share spatio-temporal datasets such as air quality but also others like meteorology or road traffic. This availability enables researchers to use these datasets in order to conduct various studies, such as the relationship between air pollution, natural phenomena (e.g., propagation by

T. Hara and H. Yamaguchi (Eds.): MobiQuitous 2021, LNICST 419, pp. 406–421, 2022.
https://doi.org/10.1007/978-3-030-94822-1_22

wind) and human activities. Many novel methods have been proposed that take not one (unimodal) but multiple inputs (multi-modal) – coming from different type of sources – in order to manage spatio-temporal datasets.

However, these methods are usually scenario-specific and few of them include functions or blocks designed to capture relationships between the different inputted modalities (cross-modal relationships). In other words, each modality is used independently and simply concatenated with the others in order to produce final outputs. However, such a technique is preventing neural networks from capturing any existing relation between the modalities. This associated loss of valuable information will prevent models in reaching more accurate performances. Especially, increasing the number and the variety of modalities will probably not allow models to capture more in-depth analysis and achieve significant improvements. For example, the performance in air quality prediction will not improve even though other modalities which would clearly affect air quality, such as human activity data, are added. This paper aims to investigate the aforementioned problem by tackling the following three questions:

1. Bui et al. said [2] that usual neural networks such as RNN do not work well when input data becomes multimodal. As a consequence, complex networks are necessary to handle multimodal datasets. Is this affirmation data-, scenario-specific or valid for any scenario?
2. Are complex architecture proposals scenario-, data-specific or are they generic enough to maintain outstanding performance with any scenario?
3. Do these complex architectures still perform well when more modalities are included? Or should they require dedicated functions or blocks to capture cross-modal relationships?

In Sect. 3, a multimodal dataset composed of air quality, meteorology and human dynamics data will be presented. This dataset is then used to answer our interrogations using different experiments (described in Sect. 4). Before concluding, Sect. 5 details the results and Sect. 6 further discusses how cross-modal relationships could be realized in order to improve future models.

Our main contributions in this paper are the following:

- We unveil a problem with recent spatio-temporal architectures. They lack significant improvement when dealing with more modalities, even though these new modalities have a clear relation with targeted data.
- We clarify the performance and robustness of one of the latest spatio-temporal proposals using a dataset with different spatial and temporal characteristics.

2 Related Research

In this section, we have selected four recent Machine Learning initiatives that deal with multimodal spatio-temporal scenario.

2.1 Multi-view Spatio-temporal Network for Taxi Demand Prediction

Yao et al. proposed a multi-view spatio-temporal network, called DMVST-Net [1], for taxi demand prediction. Their proposal method consists of 3 views, spatial-, temporal-, and semantic-view. Their architecture is composed of a local CNN to capture spatial relationships among regions; a LSTM layer to grasp temporal dependency; and embeddings in a weighted graph to encode similar taxi demand patterns among regions. Nodes of the graph represent regions, while weights represent functional similarity of demand patterns between two nodes. Each function respectively corresponds to one view (spatial, temporal, and semantic). For their experiment, they use a multimodal spatio-temporal dataset that includes taxi request and meteorological data. In addition, they consider some context information (presence of holiday, longitude and latitude of the region, etc.). However, each modality is inputted into the network independently and there is no function for modeling cross-modal relationships.

2.2 Multimodal Spatio-temporal Network Based on Encoder-Decoder Framework for Air Quality Prediction

Bui et al. proposed a multimodal spatio-temporal network, named STAR [2], targeting air quality prediction. They use a multimodal spatio-temporal dataset that includes air quality and meteorological data. As with DMVST-Net, it includes date information (presence of holiday, month, hour). Their architecture is based on an Encoder-Decoder framework. The encoder is a combination of CNN and LSTM, along with attention layers. Both air quality and meteorological data are transformed into heat-maps and fed into a CNN. Then, outputs are passed into a LSTM with weighted attention layer. Meanwhile, date information and meteorological data of the target area, along with date information, meteorological and air quality data from neighbor regions are concatenated and fed into other LSTM units. From there on, a fusion network determined best weights on hidden vectors, which are outputs of LSTMs. The decoder comprises a simple CNN-LSTM network associated with an up-sampling unit. This last unit generates a heat-map of future air quality as prediction. Similar to the method mentioned in Sect. 2.1, there is no function for explicitly modeling cross-modal relationships.

2.3 Deep Distributed Fusion Network for Air Quality Prediction

Yi et al. proposed a deep distributed fusion network, called DeepAir [3], aiming at predicting air quality. DeepAir operates in four steps. First, it performs a spatial transformation of air quality data. Then it determines embeddings of features, including meteorology and air quality data. Resulting outputs are then passed to multiple networks named FusionNet before going through a weighted merge function. Each FusionNet comprises 3 layers, a Residual Fully Connected layer that is sandwiched between 2 Fully Connected layers. Several combinations of features are fed into each FusionNet, and weighted merge function determines best weights on outputs of the FusionNet. This proposal mainly uses air quality and meteorological data, but do not explicitly focus on cross-modal relationships, even though they are concatenated and fed into FusionNet.

2.4 Encoding-Forecasting ConvLSTM Network for Air Quality Interpolation and Prediction

Le et al. applied Encoding-Forecasting ConvLSTM [4] to air quality interpolation and prediction tasks [5]. Encoding-Forecasting ConvLSTM was first proposed for precipitation nowcasting. It consists of 4 ConvLSTM layers. Both encoding and forecasting networks have 2 ConvLSTM layers. The last state of encoding network is copied as the initial states and outputs of the forecasting network. This work is very interesting in the sense that they use 4 types of spatio-temporal data: air quality, meteorological, traffic volume and driving speed. However, looking at the results, their proposal does not seem to be enough for capturing cross-modal relationships. In fact, the prediction performance with only air quality and meteorological data surpasses the performance when they use all data. Intuitively, traffic volume and driving speed data are useful for predicting air quality considering emissions of vehicles. Therefore, one would expect the performance to improve when these data are included if their proposed architecture would effectively capture associated relationships.

In summary, it is unclear if existing methods can handle many modalities properly. There is only one work which handles over 3 modalities and this work does not seem to draw the predictive power of multiples modalities.

3 Datasets

This section described the datasets used in the following experiments. Three types of spatio-temporal data have been used: air quality, meteorological and human dynamics. Air quality and meteorological comes from public Japanese datasets, while human dynamics is owned by a Japanese Mobile Network Operator (MNO).

Air Quality and Meteorological Data

Air Quality (AQ) and Meteorological (M) data from two urban areas in Japan, Tokyo prefecture and Kawasaki, have been collected from their government websites [6, 7]. As displayed in Fig. 1, the target areas cover most of Japan's central urban area and it accounts for about 12% of Japan's total population. This dataset is composed of records collected from October 1st 2017 to December 31st 2019 (2 years and 3 months) and originating from 51 monitoring stations. Every station records both hourly air quality and hourly meteorology information.

Regarding air quality, each station monitored a set of air pollutants among 12 different types, such as PM2.5, PM10, NO and NO2. PM2.5 is the only one available in all considered stations and with the least quantity of missing data. Therefore, our experiments will focus on determining the concentration of this pollutant in the air.

As for meteorological data, each station is monitoring four weather information: temperature, humidity, wind direction and wind speed. In our experiments, we use all available weather information. However, wind speed and wind direction are not used as-is. Indeed, wind information is transformed into horizontal (longitude) and vertical (latitude) components. These components are then multiplied with wind speed in order to obtain "horizontal wind speed" and "vertical wind speed". The reason behind this

preprocessing is to avoid transposing wind direction as a linearly independent value by treating it as a categorical feature.

Finally, we applied a linear interpolation to fill-up missing information within both air quality and meteorological dataset. As a result, we end up with approximately 1 million records in the dataset.

Fig. 1. Demarcation of the target areas by orange lines, Tokyo prefecture and Kawasaki city, (approximatively 2,300 km^2 area and 15 million people). © OpenStreetMap contributors

Human Dynamics Data

An MNO in Japan, named KDDI Corp., collects GPS logs of their users (several millions) who gave explicit permission to share their locations. This location information is then statistically processed in order to produce Human Dynamics (HD) data. It consists of estimating users' activity states - staying in a given location or moving - based on the spatial distribution of their locations. After that, numbers of both unique staying users and unique moving users are aggregated at each timestep and for each cell-grid defined by standards. The number of users who allow to hand out their locations represents only a subset of the total population. As a consequence, we normalized aggregated counts based on the ratio between the number of users and the total population in Japan in order to get the final data.

In this paper, we use Human Dynamics data from October 1st 2017 to December 31st 2019 (same period as previous datasets) for the target areas, Tokyo prefecture and Kawasaki. The size of a cell is 250 m by 250 m. Unique staying and moving users are counted on an hourly basis (same granularity that air quality and meteorology data).

Figure 2 illustrates profiles of Human Dynamics data. Top [resp. bottom] plot represents weekly trend of staying (Blue) and moving (Orange) users of a cell covering an office [resp. residential] area. These trends are calculated by averaging 4 weeks during the winter period. This figure shows two completely different trends. In the office

area, the number of people rapidly increases in daytime and rapidly decreases in the evening – corresponding to usual working hours. Additionally, these numbers are less important during weekend compared to weekdays. On the other hand, in the residential area, the number of people moving and staying are significantly lower than in the office area. Besides, the number of staying people decreases in daytime and increases at night during weekdays. These plots show that Human Dynamics data accurately captures commonly known characteristics of these areas.

For sake of simplicity, in the rest of the paper, we will denote Air Quality data as "*AQ*" or "*AQ* data", Meteorological data as "*M*" or "*M* data" and Human Dynamics data as "*HD*" or "*HD* data".

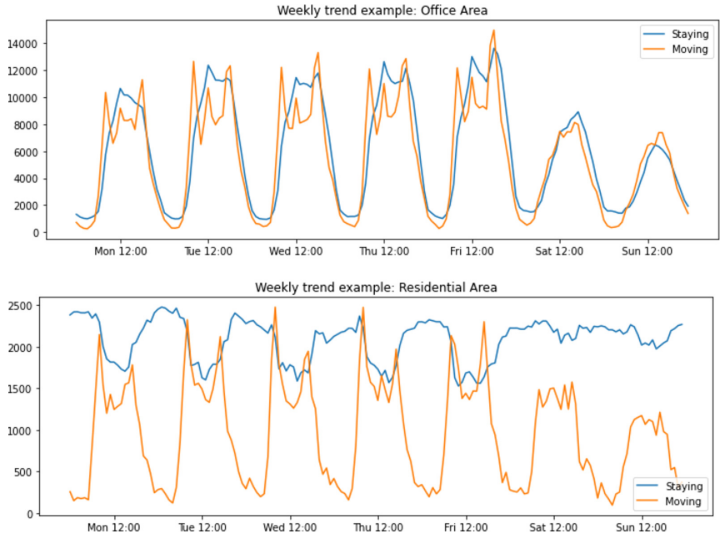

Fig. 2. Weekly trend plots of Human Dynamics data in two areas (office, top figure, and residential, bottom figure). X-axis represents the time while y-axis represents the normalized number of unique users in the area. (Color figure online)

4 Experiments

This section details the experiments that have been conducted in order to answer the three questions mentioned in Sect. 1. As most settings are the same among all experiments, we will first introduce them.

4.1 Settings

The problem setting of this study is to forecast future concentration of PM2.5 [$\mu g/m^3$] for each station. These forecasts will be performed for different prediction horizons, i.e., 1, 3, 6 and 12 h after the latest measurements' timeslot. These predictions will all be made using the past 24 h of inputs data. For example, with a prediction horizon of

3 h, when features from 07/16/2019 15:00–16:00 to 07/17/2019 14:00–15:00 are used as inputs, the corresponding output will be the concentration of PM2.5 for 07/17/2019 at 17:00–18:00. Weights are learned for each timeslot to forecast.

All features are preprocessed and transformed to [0,1] using min-max normalization.

We use the data from 10/01/2017 to 09/30/2018 as training dataset, the data from 10/01/2018 to 06/30/2019 as validation dataset and the data from 07/01/2019 to 12/31/2019 as testing dataset. The ratio of training, validation and testing is 4:3:2.

Root Mean Squared Error (RMSE) is used to evaluate prediction performance of each model, which is defined as follows:

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

where y_i and $\hat{y}_i$ represent ground truth and prediction value respectively, and where n is the number of samples (predictions). Mean Squared Error (MSE) is selected as the loss function of all models because minimizing RMSE and minimizing MSE are mathematically equivalent. We evaluate predictions and ground truths of all stations with the same weight. Models' implementation has been realized with LightGBM [8] and Keras [9] frameworks.

4.2 Experiment 1: Prediction Using Usual Model with Either Unimodal or Multimodal Data

This first experiment aims at determining if usual neural networks such as GRU or LSTM work well with multimodal inputs data. As mentioned previously, according to study [2], in air quality prediction task, prediction errors of RNN and 1D-CNN increased when meteorological features were added to historical air quality features. This result shows that simple models are probably not good at handling multimodal spatio-temporal data. However, we cannot rule out the possibility that this result specifically depends on the data they used. Therefore, the experiment described below aims to confirm whether usual models have difficulty predicting data in a multimodal configuration.

For this goal, we compare the prediction performance of GRU and LSTM when (i) using only *AQ* data (unimodal) and (ii) using both *AQ* and *M* data (multimodal). The reason for choosing GRU and LSTM as usual models is that, even if they are more complex than RNN, they remain simpler compared to the most recent proposal such as the ones presented in Sect. 2. The parameters used for this experiment are listed in Table 1 below.

4.3 Experiment 2: Prediction with Multimodal Data Using Recent Spatio-temporal Proposal and Usual Models as Baselines

With this second experiment, we try to unveil if recent spatio-temporal architecture models are generic enough to provide good prediction accuracy with any dataset, while outperforming usual models. In this experiment, we choose DMVST-Net [1] as the recent method and historical average (HA), LightGBM, GRU as well as LSTM as simple

models (also referred as baselines). Among all related researches presented in Sect. 2, DMVST-Net is the only method providing public source code. This access to the code ensures reproducibility of their results. As a result, DMVST-Net has been selected as the candidate for the recent spatio-temporal model. Although DMVST-Net originally focused on predicting taxi demand, its flexible structure makes it easy to apply it to air quality prediction. Moreover, this method was also adopted as a baseline in another study [3], which also targets air quality prediction.

Some additional preprocessing is necessary in order to apply this method to our dataset. In fact, DMVST-Net requires grid-like data as input to the Local-CNN. As air quality is available per station, we need to apply a spatial transformation in order to make it grid-like. First, we map every station to a 1 km $\times$ 1 km cell according to their latitude and longitude. Then, as inspired by the study [3], we interpolate data for cells that have no stations with Inverse Distance Weighting (IDW) [10] and we set the number of neighbor stations in IDW to 5. In our scenario, no cell contains multiple stations. As a consequence, the value of air quality of cells containing a station, solely corresponds to the value of the corresponding station. For predicting air quality of a specific station, the values of the $9 \times 9 \times 1$ cell-grid centered on the station are fed into Local-CNN. As mentioned previously, we only use one pollution data (PM2.5), that is the reason why the channel of Local-CNN's input is 1. As for meteorological data, the processing is similar to the original study; they are fed into LSTM and then, these outputs are concatenated with the output of the Local-CNN. With regard to the embeddings of stations, we first calculate each station's weekly air quality pattern by averaging from their records over the training dataset. Then, we define a weighted graph, where stations are set as nodes and weights of edges are similarities between air quality patterns of two stations. As done in the original study [1], we define similarities between patterns with the formula below:

$$S_{ij} = \exp(-DTW(i, j))$$

where i, j represents nodes, and $DTW(i, j)$ is the distance of Dynamic Time Warping (DTW) between patterns. After that, we apply LINE [11] for generating embeddings of nodes. The dimension of embeddings is set to 32, same as in the original study.

However, contrary to the original study and because of GPU memory limitations, we have reduced by half some parameters in DMVST-Net (number of filters in the Local-CNN and hidden dimension of the LSTM layer). We use Adam optimizer and parameters beta_1, beta_2, epsilon and decay are set to the same values as the original study. The other parameters are summarized in Table 1.

We describe settings of baselines below:

Historical Average: Average of the last 24 h is used as prediction.

LightGBM: We use the latest values and some aggregated values in last 24 h of concentration of PM2.5, temperature, humidity, horizontal and vertical components of wind. As aggregated values, we calculate max/min/average values in the last 3/6/12/24 h. Thus, the number of features is of $5 + (5 * 3 * 4) = 65$. We do not normalize feature values and target values. After generating features, we search some hyperparameters for

validation set by Bayesian Optimization. Target hyperparameters are 'num_leaves', 'feature_fraction', 'bagging_fraction', 'min_data_in_leaf' and 'min_sum_hessian_in_leaf'. We set 'init_points' to 5 and tuning iterations to 5. Iterations of LightGBM itself are set to 10000.

The parameters of LSTM and GRU used for this experiment are listed in Table 1.

4.4 Experiment 3: Prediction with Multimodal Data Using a Modified Version of Recent Spatio-temporal Proposal

This third experiment seeks to determine if recent spatio-temporal proposal can maintain good performances with more modalities. For this purpose, we extended DMVST-Net (referred as Ex-DMVST-Net) so that it can support one more modality. Then, we compare prediction performances between the original DMVST-Net that uses *AQ* and *M* data and Ex-DMVST-Net that uses *AQ*, *M* and *HD* data.

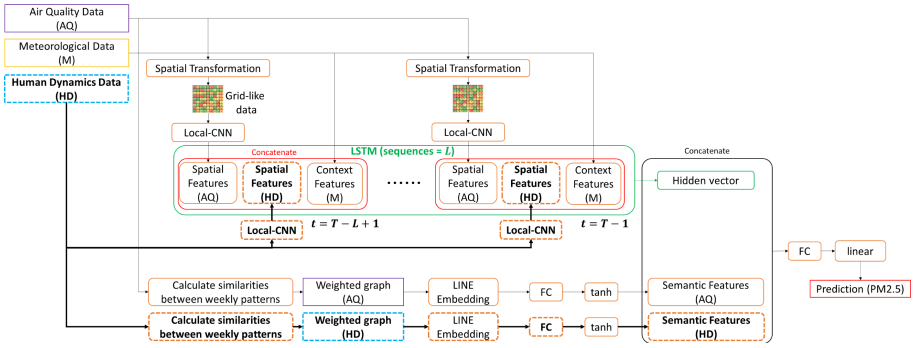

Fig. 3. The architecture of Ex-DMVST-Net. Blocks with bold-type letters and dashed border represent extended parts from original DMVST-Net.

The architecture of Ex-DMVST-Net, and changes made from the original DMVST-Net, is shown in the Fig. 3. In this figure, blocks that are drawn with bold-type letters and dashed border are our extensions for including *HD* as input from the original DMVST-Net. In Ex-DMVST, *HD* is treated like *AQ*. It is fed into a new Local-CNN block, then the output is concatenated with other features, i.e., the output of AQ Local-CNN and meteorological features generated from *M*. The resulting concatenation is then inputted into the LSTM layer. Meanwhile, *HD* is also used to generate a weighted graph of human dynamics by calculating similarities between HD patterns of two cells.

In the same manner as *AQ*, this weighted graph is fed into LINE and embeddings of cells are generated. These embeddings go through a fully-connected layer followed by a *tanh* activation and become Semantic Features of *HD*.

HD input is a cell-grid of shape $17 \times 17 \times 2$ and centered on the target station. As aforementioned, *HD* has 2 types of values, number of staying and moving people, therefore channel of *HD* input is 2. These inputs are then fed into a dedicated Local-CNN. Before generating the weighted graph, *HD* data is divided into 2×2 parts, staying/moving

and weekday/holiday combinations. As shown in Fig. 2, weekdays patterns are very similar in most cells (and it is also the case for weekends patterns).

As a result, we first generate 4 daily patterns for each cell, then concatenate and treat them as a 96-dimensional vector (24 h $\times$ 4 patterns). We calculate daily patterns from the training set but exclude days with specific events or holidays (e.g., early May, mid-August and around New Year are holiday season in Japan). As a matter of fact, *HD* patterns during days with specific events/holidays differ completely from usual weeks. Contrary to AQ, in *HD*, Pearson correlation is used as the similarity method between patterns.

As Ex-DMVST-Net receives one more modality, a much larger size of GPU memory has to be allocated. As a consequence, the batch size is set to 8 * 51 and the learning rate is to 2.5e−5 in order to cope with our GPU limitations. All the other parameters are exactly the same as the original model, as summarized in Table 1.

Table 1. Models' parameters

Method	GRU		LSTM		DMVST-Net	Ex-DMVST-Net
Inputs	AQ	AQ + M	AQ	AQ + M	AQ + M	AQ + M + HD
Hidden dimension	128	128	128	128	LSTM 256 Local-CNN 16	LSTM 256 Local-CNN 16
Number of layers	1	1	1	1	1/3	1/3
Batch size	128 * 51	128 * 51	128 * 51	128 * 51	128 * 51	8 * 51
Learning rate	1.0e−2	5.0e−3	1.0e−2	1.0e−2	1.0e−4	2.5e−5
Optimizer	SGD	SGD	SGD	SGD	Adam	Adam
Momentum	0.9	0.9	0.9	0.9		
Beta_1, beta_2, epsilon, decay					0.9, 0.999, 1e−08, 1e−6	0.9, 0.999, 1e−08, 1e−6
Early stopping (patience)	Yes (10)	Yes (10)	Yes (10)	Yes (10)	Yes (10)	Yes (10)
Number of epochs	100	100	100	100	100	100
Batch normalization	No	No	No	No	No	No
Dropout	No	No	No	No	No	No

5 Results

This section presents the results of experiments introduced in Sect. 4.

5.1 Prediction Using Usual Model with Either Unimodal or Multimodal Input

As described in Sect. 4.2, performances of both GRU and LSTM have been tested when using AQ only (unimodal) and $AQ + M$ (multimodal). Table 2 lists the result of this experiment. As presented in Sect. 4.1, for each model AQ predictions will be computed at four different time horizons. Regarding GRU, adding one more modality is improving accuracy for two prediction horizons. On the contrary, an additional modality with LSTM is only improving RMSE for one out of four prediction horizons. These results confirmed the claims from previous study [2] that simple models are not good at handling multimodal spatio-temporal data.

Table 2. Prediction errors (RMSE) of GRU and LSTM with different inputs

Method	Inputs	+1 h	+3 h	+6 h	+12 h
GRU	AQ	3.265	**4.640**	5.807	**6.615**
GRU	$AQ + M$	**3.160**	5.003	**5.477**	7.348
LSTM	AQ	3.222	**4.689**	**5.774**	**6.621**
LSTM	$AQ + M$	**3.144**	4.890	6.895	7.178

5.2 Prediction with Multimodal Data Using Recent Spatio-temporal Proposal and Usual Models as Baselines

As mentioned in Sect. 4.3, prediction errors (RMSE) of DMVST-Net will be compared to various baselines, as presented in Table 3. $AQ + M$ is used as inputs for all methods. As shown in the table, DMVST-Net achieves better performance than historical average, GRU and LSTM. However, LightGBM surprisingly shows the best performance in 3 out of 4 predictions horizons. One probable reason for this interesting result is that only LightGBM uses aggregated features – min/max/average. They may capture general trends in target that help LightGBM better comprehend data. Therefore, DMVST-Net might achieve better performance for all prediction settings if aggregated features were fed into DMVST-Net, similar to LightGBM.

Table 3. Prediction errors (RMSE) of DMVST-Net and baselines with $AQ + M$ inputs

Method	+1 h	+3 h	+6 h	+12 h
HA	5.899	6.319	6.803	7.464
LightGBM	2.989	**4.398**	**5.387**	**6.282**
GRU	3.160	5.003	5.477	7.348
LSTM	3.144	4.890	6.895	7.178
DMVST-Net	**2.966**	4.521	5.756	6.645

5.3 Prediction with Multimodal Data Using a Modified Version of Recent Spatio-temporal Proposal

As shown in Table 4, this experiment compares prediction errors (RMSE) obtained with DMVST-Net and our extended version of DMVST-Net (Ex-DMVST-Net). Similar to Experiment 2, DMVST-Net is fed with $AQ + M$, while $AQ + M + HD$ datasets are fed into Ex-DMVST-Net. Prediction error increased in 3 out of 4 prediction settings when *HD* is added, and the improvement carried out in the last setting is not significant (~0.3%). This result indicates that Ex-DMVST-Net cannot utilize the potential of *HD* when predicting air quality.

Table 4. Prediction errors (RMSE) of DMVST-Net and Ex-DMVST-Net with different inputs

Method	Inputs	+1 h	+3 h	+6 h	+12 h
DMVST-Net	$AQ + M$	**2.966**	**4.521**	**5.756**	6.645
Ex-DMVST-Net	$AQ + M + HD$	3.210	4.684	5.779	**6.627**

6 Discussion

The previous experiments show that simply adding new modalities to prediction models does not guarantee an improvement in accuracy. The first experiment confirms that usual models are not handling properly multiple modalities. These results extend the study [2] and demonstrate that this phenomenon is model-specific and does not depend on the datasets. With more recent architectures, as shown by DMVST-Net in Experiment 2, it is possible to achieve better results when using multiple modalities. However, even if new architecture principles are extended to handle more modalities, it does not improve results, as shown in the third experiment. These experiments suggest that proposed models are not generic enough in order to handle any number of multiple modalities. As a consequence, there is a need for a generic mechanism that would enable a given architecture to achieve significant performance improvement using any number of modalities. We believe that this mechanism – referred in the rest of the paper as the *relationship mechanism* – would require to learn about both modality relationship with the target and cross-modal relationships.

Nevertheless, one could argue that performance may vary according to the modalities selected. In fact, in the case of the third experiment, there is no proof that HD data have the said potential to improve AQ predictions. Due to the relatively new availability of Human Dynamics data, no reference has been found to back-up this intuition.

We have conducted an additional study in order to investigate correlation between HD and AQ. Considering the strong relation of HD data with spatial position and time, it is difficult to determine a unique Pearson correlation value with AQ, which is defined per station. Indeed, correlation may vary with each cell. In addition, we assumed that HD values of cells directly surrounding a station will influence AQ measurements; but it is difficult to estimate the radius of surrounding cells as it might also depend on other parameters, such as wind or environment.

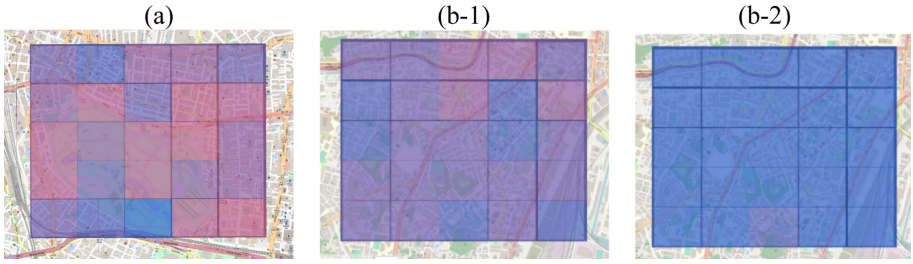

Fig. 4. Examples of correlation levels between the number of moving people and AQ for a 5 × 5 grid cells centered on different stations. (a) An example of correlation levels around a given station for any type of day and time. (b) Examples of correlation levels around another station on weekdays during (b-1) morning commuting period (5:00 to 11:00) and (b-2) late at night (1:00 to 5:00). (Color figure online)

Figure 4 shows examples of correlation levels between HD (the number of moving people) and AQ for grid-cell centered on two different monitoring stations. Red and blue shading represents the range of correlation from high to low respectively. As expected, and shown in Fig. 4a, depending on the location, correlations greatly vary. Nonetheless, sub-Figs. 4b unveil an interesting finding. In fact, HD correlation with air pollution of cells that cover main roads are greater than cells without. This result correlates with the fact that cars emit exhaust gas that directly impact air quality. Besides, Human Dynamics represents human behavior based on their position at regular time-intervals. Sub-Figs. 4b-1 and 4b-2 illustrate the correlation between HD and AQ for the same area, on weekdays, but for different time period (from 5 am to 11 am and 1 am to 5 am from respectively). These sub-figures clearly demonstrate that during commuting peak period (i.e., 5 am to 11 am) correlation with air pollution is higher for cells with main roads. On the other hand, late at night (i.e., 1 am to 5 am) correlations are almost the same for all considered cells of the grid. Same finding, but with different level of correlation, can be noticed for other stations. Time is playing an important role in this correlation study. And as mentioned previously, other parameters may influence correlation, such as the topography. These parameters could induce delays in the impact of HD on AQ, especially for cells that are relatively far from the station. Figure 5 plots distribution for all the considered stations of lagged Pearson correlation between HD and AQ. It shows that adding lag to the number of staying people has no impact on the studied correlation, while the number of moving people may have different delays with different stations. Indeed, the distribution spread more on the positive side (i.e., right) when adding a delay compared with no delay. Further analysis will be required to fully explain these results.

This investigation confirms that a correlation exists between AQ and both the number of moving and staying people. However, it is not straightforward as it depends on location, time and some delays. In addition, it appears that the definition of HD might be too generic. In fact, the number of moving people aggregates all the persons moving in a given cell. However, each transport mode will not have the same effect on AQ. Current HD data are mixing means of transportation that have no impact on AQ (e.g., walking, cycling, etc.) with those that have a negative impact on AQ (e.g., driving, etc.). It is therefore of crucial importance to have HD datasets that separate moving people based

on their transport modes. Such new datasets associated with the *relationship mechanism* should help model better predict air pollution.

Finally, adding modalities to model is increasing the complexity for the model to find prediction patterns and thus does not guarantee better performance in terms of accuracy. In order to fully benefit from multiple modalities, it is important to have a specific mechanism to learn intra- and inter-relationships of inputs data. We are planning to further investigate this idea by adding a cross-modal attention block to the architecture. Indeed, adding spatial and temporal attention independently for each modality may not be sufficient. An additional block might be necessary to capture the cross-modal combination to determine how modalities affect each other.

In addition, meteorological data is unique compared to other datasets in our scenario. In fact, it may dynamically change relationships of other modalities. For example, the HD data of cells that may affect the PM2.5 concentration level of a given cell may completely vary according to the wind information. As proposed in a recent paper [12], modeling the propagation of PM2.5 could help models better predict air pollution. Unfortunately, this proposal does not support more than two modalities. To solve this problem, we propose to model propagation of PM2.5 in the atmosphere based on wind speed and direction and combine it with cross-modal effect of multiple modalities.

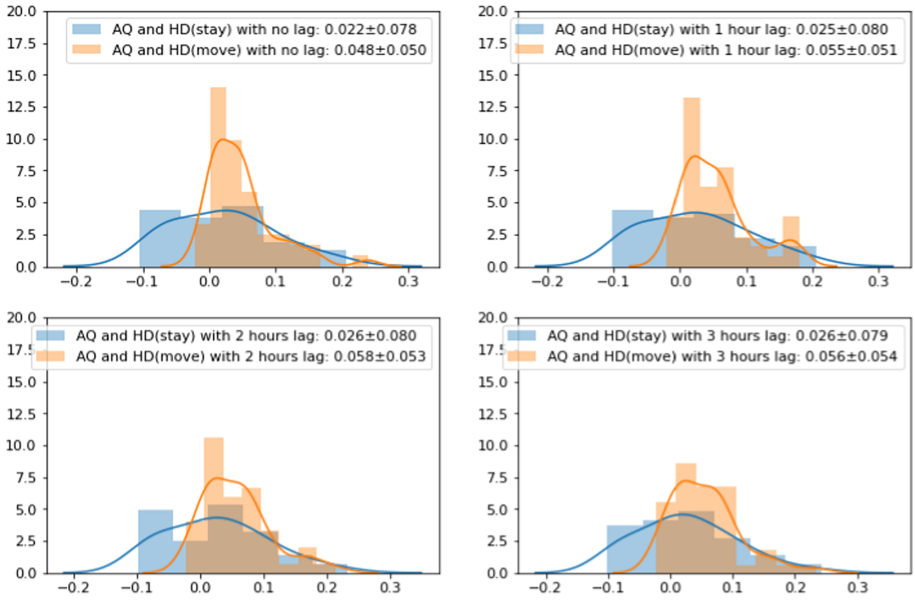

Fig. 5. Distributions for all the considered stations of lagged Pearson correlation between AQ and HD (people moving and staying). Lags are set from 0 to 3 h.

7 Conclusion

In this paper, we unveil a common and yet significant problem among studies which focus on handling multimodal spatio-temporal data. Indeed, adding more modalities does not guarantee an improvement in prediction accuracy, even though these additional modalities are related to the predicted data. We conduct some experiments in order to confirm that the problem exists with a recent spatio-temporal architecture proposal. The study shows that, even though efficient for Taxi Demand, the architecture DMVST-Net does not provide satisfying performance with Japan Air Quality dataset. And adding information on human activity, that correlates with air quality to some extent, is not improving the results. This study confirms the assumptions that recent Deep Learning architecture models might not be able to capture complex cross-modal relationships in any scenario.

For future work, we plan to propose a novel model architecture that can handle cross-modal relationships even though many modalities are inputted. In order to realize this, we advance to use specific attention layers to capture cross-modal relationships. And to further improve the knowledge of models in air pollution prediction scenario, we suggest to combine the former proposal with an explicit propagation model of air pollution in the atmosphere. Finally, for more practicality, we will consider providing a way to interpret these relationships, such as a visualization of attention weights.

Acknowledgments. This work was supported by KDDI Corporation in providing Human Dynamics Data.

References

1. Yao, H., et al.: Deep multi-view spatial-temporal network for taxi demand prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1 (2018)
2. Bui, T.-C., et al.: STAR: spatio-temporal prediction of air quality using a multimodal approach. In: Arai, K., Kapoor, S., Bhatia, R. (eds.) IntelliSys 2020. AISC, vol. 1251, pp. 389–406. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-55187-2_31
3. Yi, X., Zhang, J., Wang, Z., Li, T., Zheng, Y.: Deep distributed fusion network for air quality prediction. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 965–973. Association for Computing Machinery, New York (2018)
4. Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., Woo, W.: Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Proceedings of the Advances in Neural Information Processing Systems, vol. 28 (2015)
5. Le, V., Bui, T., Cha, S.: Spatiotemporal deep learning model for citywide air pollution interpolation and prediction. In: 2020 IEEE International Conference on Big Data and Smart Computing, pp. 55–62 (2020)
6. Bureau of Environment, Tokyo Metropolitan Government. https://www.kankyo.metro.tokyo.lg.jp/air/air_pollution/torikumi/result_measurement.html. Accessed 17 June 2021
7. Kawasaki City Website. https://www.city.kawasaki.jp/kurashi/category/29-1-10-2-1-7-0-0-0-0.html. Accessed 17 June 2021
8. LightGBM. https://lightgbm.readthedocs.io/en/latest/. Accessed 18 June 2021

9. Keras. https://github.com/fchollet/keras. Accessed 18 June 2021
10. Lu, G.Y., Wong, D.W.: An adaptive inverse-distance weighting spatial interpolation technique. Comput. Geosci. **34**, 1044–1055 (2008)
11. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: LINE: large-scale information network embedding. In: Proceedings of the 24th International Conference on World Wide Web, pp. 1067–1077. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2015)
12. Wang, S., Li, Y., Zhang, J., Meng, Q., Meng, L., Gao, F.: PM2.5-GNN: a domain knowledge enhanced graph neural network for PM2.5 forecasting. In: Proceedings of the 28th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '20), pp.163–166. Association for Computing Machinery, New York (2020)