# Conditional Variational Graph Autoencoder for Air Quality Forecasting

*Esther Rodrigo Bonet*[⋆‡], *Tien Huu Do*[⋆‡], *Xuening Qin*[†‡], *Jelle Hofman*[‡], *Valerio Panzica La Manna*[‡], *Wilfried Philips*[†‡], *Nikos Deligiannis*[⋆‡]

[⋆] ETRO Department, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium
[†] IPI, Ghent University, Sint-Pietersnieuwstraat 25, B-9000 Ghent, Belgium
[‡] imec, Kapeldreef 75, B-3001 Leuven, Belgium

***Abstract*—To control air pollution and mitigate its negative effect on health, it is of the utmost importance to have accurate real-time forecasting models. Existing deep-learning-based air quality forecasting models typically deploy temporal and—less often—spatial modules. Yet, data scarcity emerges as a real issue in this domain, a problem that can be solved by capturing the data distribution. In this work, we address data scarcity by proposing a novel conditional variational graph autoencoder. Our model is able to forecast air pollution by efficiently encoding the spatio-temporal correlations of the known data. Additionally, we leverage dynamic context data such as weather or satellite images to condition the model's behaviour. We formulate the problem as a context-aware graph-based matrix completion task and utilize street-level data from mobile stations. Experiments on real-world air quality datasets show the improved performance of our model with respect to state-of-the-art approaches.**

***Index Terms*—Air quality forecasting, conditional variational graph autoencoders, context-aware graph-based matrix completion, deep learning.**

## I. Introduction

Air pollution has become a world-wide issue in the last decades, hence the need for accurate air quality forecasting solutions. To measure pollutant concentrations in urban areas, mobile and fixed monitoring stations have been effectively deployed. Fixed stations can collect measurements with high temporal resolution, yet, they are expensive, hence their low spatial resolution. On the contrary, mobile sensors allow for high spatial resolution, however, their temporal resolution per location is low since the sensors are mounted on moving vehicles. Air pollution inference has been the preferred approach to solve the issue of data scarcity due to missing measurements and the uneven distribution of the observed data. Nevertheless, inference is not sufficient if we wish to mitigate air pollution effects, where forecasting might be of higher relevancy. This renders spatio-temporal forecasting of air quality a highly interesting task.

Various approaches have been proposed to predict future pollutant concentrations *using measurements coming from ground-level stations*, including shallow methods [1]–[6] and deep-learning approaches [7]–[9]. Shallow approaches are based on physical models, statistical methods or shallow artificial neural networks, amongst others [10]. Such models require a deep understanding of complex particle dispersion mechanisms, leading to computationally prohibitive complete solutions and low performance in approximate solutions [10]. Recently, deep networks have emerged as a potential technology to extract complex features from high-dimensional pollution data [11]. These models do not rely on strong assumptions; instead, they require huge amounts of pollution data and diverse context data such as weather or traffic. As such, deep models have achieved promising performance results in air quality forecasting [12], [13].

Very limited work has focused on air quality forecasting *using data collected by mobile stations* [14]. In [13], we used mobile stations from the City-of-Things (CoT) platform [15] to tackle the task of air pollution estimation in unmeasured locations, hence solving an inference problem. Contrarily, in this paper, we focus on forecasting air pollutant measurements at certain locations and at future time instants. Forecasting air pollution measurements can be relevant for various applications in smart cities, including safeguarding public health, navigating traffic levels or optimising city planning. For the task, we employ pollution data collected from the CoT platform [15] and Kunak sensors in the Bel-Air[1] project. Unlike [13], where additional knowledge was not considered, we leverage context data to condition the model. We follow a data-driven approach and formulate the air quality forecasting problem as a context-aware graph-based matrix completion problem. Specifically, we propose a novel deep learning model based on variational graph autoencoders conditioned to the context data; we refer to our model as cVGAE. The model is conditioned by other data types such as weather, points of interest (POIs) or satellite images and effectively captures the spatio-temporal dependencies in the measurements. Experiments on real data show that our method outperforms various reference models.

To summarize, the main contributions in our paper are: (i) we formulate air quality forecasting as a graph-based matrix completion problem and propose a variational graph autoencoder with condition for accurate forecasting. To the

This work was supported in part by the Research Foundation - Flanders (FWO) through the Ph.D. Fellowship Strategic Basic Research under Project 1SC4521N, in part by IMEC under the AAA Project *AI-based Air Quality Map and Analytics* and in part by the Flemish Government, under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

[1] https://www.imeccityofthings.be/en/projecten/bel-air

best of our knowledge, this is the first work to leverage conditional variational graph autoencoders to perform air quality forecasting; (ii) the proposed architecture effectively learns the spatial and temporal correlations of the air pollution and context data via graph convolutional operations and by imposing temporal and spatial smoothness constraints; (iii) we present comprehensive experiments on real-world datasets and achieve superior performance compared to other models.

The remainder of the paper is organised as follows. Section II reviews related work. Section III presents the problem and the proposed forecasting model. Section IV describes the experimental results and Section V concludes our work.

## II. Related Work

### A. Air Pollution Forecasting

Air pollution forecasting has been addressed by deploying physical models [1], statistical methods [2], [3] and shallow artificial neural networks (ANNs) [4]–[6]. Such approaches, however, require specific domain knowledge and huge computational power or do not capture the latent correlations of air quality data with alternative data such as weather, satellite images and points of interest (POI). Recently, deep learning (DL) has matched the performance of traditional shallow methods [16]. Spatio-temporal deep learning (STDL) models, for instance, have proven their accuracy in air quality forecasting [10]. Methods in this category independently couple temporal and spatial modules in a single architecture, sometimes by incorporating context information [7], [9]—from weather, traffic, etc—or by imposing objective constraints [8]. Our work belongs to the STDL category but differs from existing methods in that we utilize a conditional variational graph autoencoder, which allows to jointly learn the latent spatio-temporal distribution of the known historical data while conditioning the model to additional contextual data. In the experiments section, we compare the performance of our method against an statistical model—namely, the autoregressive integrated moving average (ARIMA) model, an integrated version of [2]—and a STDL model—namely, the GRU+DNN [17]—and demonstrate the superior performance of our method in air quality forecasting.

### B. Conditional Variational Graph Autoencoders

Variational autoencoders (VAEs) [18] are deep generative models that have lately achieved impressive results in different domains, including air quality forecasting [19], [20], and prediction [21] or collaborative filtering [22]. Such models propose an encoder-decoder architecture with fully connected neural networks on non-structured data. Variational graph autoencoders (VGAE) [23] apply the idea behind VAEs in graph-structured data, with applications in link prediction [23] and graph generation [24]. Alternatively, conditional VAEs [25], [26] impose a condition on both encoder and decoder inputs to have control on the generative process. Conditional graph VAEs (CVGAEs) were first introduced in [27] and later in [28] to drive molecule generation. Similarly, [29] applies CVGAEs on structure-aware writing. This work is similar to ours in that it employs graph-based and constrained VAEs, however, we are able to jointly learn the correlations in single architecture rather than independent modules while aggregating additional context data. To the best of our knowledge, no previous work has tackled air quality forecasting with conditional variational graph autoencoders.

## III. Proposed Method

### A. Problem Formulation and Notation

We solve the task of air quality forecasting at the street network of urban areas; namely, we only consider street locations. For this, we are given multiple time series of pollutant concentrations collected by mobile sensor-equipped vehicles at certain locations. As the time and location associated to a measurement are continuous, the measurements are aggregated at discrete time instances and locations. We uniformly divide the time span of the data into equal slots of duration $t_D$ (e.g., one hour). In a given timeslot $t$, we gather all measurements within a predefined geographical distance $r$ from a given spatial location $\boldsymbol{y}$ on the street network and take their median-value as the measurement at location $\boldsymbol{y}$ at timeslot $t$. Hence, the aggregation across space is non-uniform and is adapted to the considered locations on the street network. We summarise the sets of time series in matrix $\boldsymbol{X} \in \mathbb{R}^{N \times T}$, with $N$ and $T$ the number of considered geographical locations and total time slots, respectively. Each row represents the time series of pollutant concentrations in a certain location $\boldsymbol{y}$ within the time span $T$; equivalently, each column represents all the measured pollutant concentrations at a certain time slot $t$. Note that the aggregation process results in matrix $\boldsymbol{X}$ being highly sparse. Let $\kappa$ and $\tau$ denote the number of past and future time instants with respect to present time $t$, that is, the past measurements are $\boldsymbol{X}_{(t-\kappa,t]} \in \mathbb{R}^{N \times \kappa}$ and the future measurements are $\boldsymbol{X}_{(t,t+\tau]} \in \mathbb{R}^{N \times \tau}$, with $T = \kappa + \tau$.

Apart from past pollutant concentrations, we collect context knowledge in the considered locations. Specifically, we process spatial features—namely, geo-coordinates, points of interest (POIs), and satellite images—and temporal features—namely, weather and the timestamp of the event. The aggregation of the context data (see Section III-B) results in the matrix $\boldsymbol{C} \in \mathbb{R}^{N \times d}$, where $d$ is the predefined dimensionality of the embedding.

In this work, we aim at forecasting the time series in $(t, t+\tau]$. By learning the latent distribution of the known past pollution data, we are able to perform forecasting at future time instances. The predicted pollutant concentrations are grouped in the matrix $\tilde{\boldsymbol{X}}_{(t,t+\tau]} \in \mathbb{R}^{N \times \tau}$.

### B. The Proposed cVGAE Model for Air Quality Forecasting

The architecture of our forecasting model, which we refer to as cVGAE, is depicted in Fig. 1. We consider the $N$ corresponding discretized locations on the street network and build a graph of $N$ nodes. Two nodes are connected if the geodesic distance between them is smaller than a predefined distance threshold $\delta$, or if they belong to the same road segment. The weight of a connection is the inverse of the

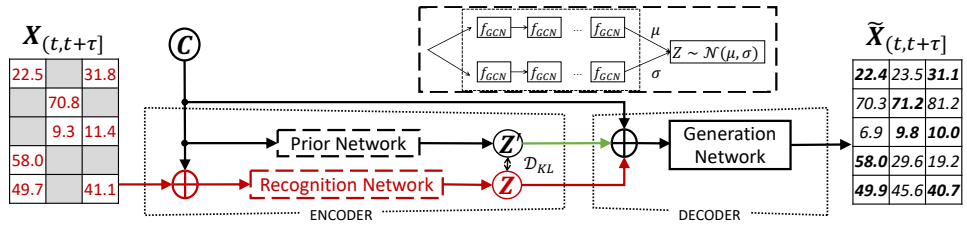


Fig. 1. The proposed conditional variational graph autoencoder for air quality forecasting (cVGAE). Elements in red font depict that these are only available during training; green elements are only activated during testing. During training, the input of the model consists of the sparse matrix $\boldsymbol{X}_{(t,t+\tau]}$ and the context matrix $\boldsymbol{C}$. $\bigoplus$ applies a concatenation operation on the matrices. Light grey cells represent unmeasured locations at certain time instants. Prior and recognition networks follow the architecture on the top-right dashed square, where the function blocks $f_{GCN}$ represent GCN layers and the variables $\mu$ and $\sigma$ describe the learnt Gaussian distribution. The output matrix $\tilde{\boldsymbol{X}}_{(t,t+\tau]}$ contains the forecasted pollutant concentrations. Bold entries are the known values on which we evaluate the loss function.

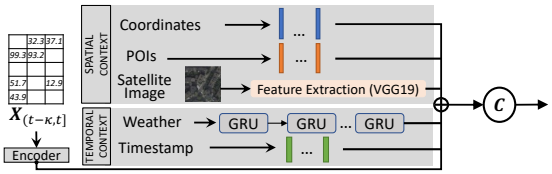


Fig. 2. The computation of the matrix of context embeddings $\boldsymbol{C}$. The spatial and temporal additional knowledge are preprocessed and concatenated with the encoded past pollution data to create a spatio-temporal embedding for each considered location. These are grouped in matrix $\boldsymbol{C} \in \mathbb{R}^{N \times d}$ which serves as a condition and/or input in the proposed method.

geodesic distance in meters computed by the Haversine formula [29]. The proposed model, which follows an encoder-decoder architecture, is based on VGAEs and conditioned by context information in $\boldsymbol{C}$.

We build the matrix of context embeddings $\boldsymbol{C}$ from past time instances i.e., within $(t-\kappa, t]$ as depicted in Fig. 2. To combine the spatial—coordinates, POIs, satellite images—and temporal—weather and timestamp—context information with the past air pollution data, we first preprocess the data to obtain latent embeddings of each source of information. These are later concatenated in a single context embedding of size $d$.

The encoder in Fig. 1 consists on two VGAEs, namely, prior and recognition networks. Their architecture, equivalent for both networks, is depicted in the dashed square of Fig. 1. The prior network learns the prior distribution $p(\boldsymbol{z}|\boldsymbol{c})$ where $\boldsymbol{z}$ and $\boldsymbol{c}$ are the latent variable and the context embedding of a certain location. Similarly, the recognition network approximates the posterior distribution of $p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c})$ where $\boldsymbol{x}$ is the pollution time series for a certain location, i.e., $\boldsymbol{x}_{(t,t+\tau]}$. We wish to maximize $p(\boldsymbol{x}|\boldsymbol{c})$, hence the need to reduce the Kullback–Leibler (KL) distance between $p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c})$ and $p(\boldsymbol{z}|\boldsymbol{c})$ in the loss function (see Section III-C). Note that during testing, we only have access to past time instances, i.e., $\boldsymbol{X}_{(t,t+\tau]}$ is unavailable. We reflect this condition by making $\boldsymbol{Z}$ unavailable during testing, i.e., the recognition network is ignored and the decoder employs $\boldsymbol{Z'}$ as input.

The decoder in Fig. 1 consists on a generation network whose input differs in the training and testing phase. During training, the generation network handles the transformation of $\boldsymbol{Z}$, i.e., the data distribution generated by the recognition network—conditioned by $\boldsymbol{C}$ the context knowledge—to the matrix $\tilde{\boldsymbol{X}}_{(t,t+\tau]}$ of air pollution forecasts. During testing, only $\boldsymbol{Z'}$ is available. In this phase, we sample from the prior distribution $\boldsymbol{Z'}$ and take as input $\boldsymbol{C}$ to forecast $\tilde{\boldsymbol{X}}_{(t,t+\tau]}$. For this task, we design two decoders; specifically, we test an architecture of stacked GCN layers—which we refer to as cVGAE in the experiments section—and an autoregressive generative model—referred as cVGAE Autoregressive—which aggregates stacked GRU layers prior to the GCN layers.

### C. The Loss Function

The loss function of our model is defined in (1). We adapt the loss function defined in [18] for a forecasting task by using the mean absolute error (MAE) and correlation error—$\ell_1$ and corr in (1), respectively—regularized by a KL divergence term. Even though the MAE is not everywhere differentiable, we find that using its sub-gradient is sufficient for optimization with gradient descent. The temporal dependency between measurements ($\ell_{\text{temp}}$) imposes an additional smoothness constraint. Additionally, we avoid model overfitting by applying $\ell_2$-regularization on the network weights ($\ell_{2,\boldsymbol{W}}$):

$$\mathcal{L} = \ell_1 + \alpha(1 - \text{corr}) + \beta\bar{\mathcal{D}}_{\text{KL}} + \lambda\ell_{2,\boldsymbol{W}} + \gamma\ell_{\text{temp}} \quad (1)$$

where $\alpha, \beta$ and $\gamma$ are positive tuning parameters and $\lambda$ is the weight decay parameter.

The $\ell_1$ norm computes the MAE between forecasted and known pollutant concentrations as in:

$$\ell_1 = \frac{1}{|\Omega|}\sum_{\Omega}|\tilde{\boldsymbol{X}}_{(t,t+\tau]} - \boldsymbol{X}_{(t,t+\tau]}|. \quad (2)$$

where $\Omega$ denotes the set of known entries on the training set.

Similarly, the correlation loss is computed between forecasted and known pollutant concentrations:

$$\text{corr} = \frac{\text{cov}(\tilde{\boldsymbol{X}}_{(t,t+\tau]}, \boldsymbol{X}_{(t,t+\tau]})}{\sigma_{\tilde{\boldsymbol{X}}_{(t,t+\tau]}} \cdot \sigma_{\boldsymbol{X}_{(t,t+\tau]}}} \quad (3)$$

where the covariance $\text{cov}(\cdot,\cdot)$ and the standard deviations $\sigma$ are computed on the elements in $\Omega$.

Additionally, we avoid model overfitting by applying $\ell_2$-regularization on the following network weights: matrices $\boldsymbol{W}_{POI}, \boldsymbol{W}_s, \boldsymbol{W}_h, \boldsymbol{W}_r$—which refer to the learnt weights,

TABLE I
DESCRIPTION OF THE COT AND KUNAK $NO_2$ DATASETS. THE UNITS ARE PARTS PER BILLION (PPB) AND $\mu g/m^3$, RESPECTIVELY.

| Dataset | CoT | Kunak |
|---|---|---|
| Number of locations | 3630 | 4953 |
| Duration in hours | 720 | 1464 |
| Max concentration | 633.65 | 111.92 |
| Min concentration | 0.16 | 0.02 |
| Mean concentration | 85.50 | 32.88 |
| % of known entries versus all | 0.60 | 0.58 |

needed to process the POI, satellite images and past pollution data and to produce the matrix of context embeddings $\boldsymbol{C}$, respectively—, matrices $\boldsymbol{W}_{\boldsymbol{\mu}'}, \boldsymbol{W}_{\boldsymbol{\sigma}'}, \boldsymbol{W}_{\boldsymbol{\mu}}, \boldsymbol{W}_{\boldsymbol{\sigma}}$—which denote the learnt weights to process the prior and recognition networks—and matrix $\boldsymbol{W}_{GCD}$—which denotes the learnt weights for the decoder—. Note that the matrices $\boldsymbol{\mu}', \boldsymbol{\sigma}'$ and $\boldsymbol{\mu}, \boldsymbol{\sigma}$ denote the means and standard deviations of the two multivariate Gaussian distributions learnt by the prior and recognition networks, i.e., $\boldsymbol{Z}' \sim \mathcal{N}(\boldsymbol{\mu}', \sigma')$ and $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$, respectively. Furthermore, the KL divergence term $\mathcal{D}_{\text{KL}}$, which can be computed with a closed form formula [18], calculates the matrix of KL distances between data distributions $\boldsymbol{Z}'$ and $\boldsymbol{Z}$. The average of $\mathcal{D}_{\text{KL}}$ is computed and incorporated in (1).

Finally, the $\ell_{\text{temp}}$ term applies a smoothing constraint on the air pollutant forecasts over time:

$$\ell_{\text{temp}} = \sum_{w=1,\ldots,w_T} e^{-w} \sum_{i,j} \left| \tilde{\boldsymbol{X}}_{(t,t+\tau]}(i,j) - \tilde{\boldsymbol{X}}_{(t,t+\tau]}(i,j+w) \right| \tag{4}$$

where $\tilde{\boldsymbol{X}}_{(t,t+\tau]}(i,j)$ is the element in the $i$-th row and $j$-th column of $\tilde{\boldsymbol{X}}_{(t,t+\tau]}$, i.e., the forecasted pollutant concentration at location $i = \{1, \ldots, N\}$ and time instant $j = \{1, \ldots, \tau - w\}$. The width of the neighborhood $w_T$ is a parameter that is fine-tuned experimentally.

We minimize the loss function in (1) with respect to the training entries using the stochastic gradient descent—where we use the reparameterization technique in [18]—and we deploy the dropout regularization technique to mitigate overfitting. After training, we obtain matrix $\tilde{\boldsymbol{X}}_{(t,t+\tau]}$ containing the pollutant concentration forecasts at future time instances.

## IV. EXPERIMENTS

### A. Description of the Datasets

We rely on the CoT platform [15] and Kunak sensors to collect two datasets with $NO_2$ air quality measurements in the city of Antwerp, Belgium. The former corresponds to measurements collected in May 2018 while the latter contains measurements collected during March and April 2021.

The preprocessing of the data consists on the aggregation step presented in Section III-A. We follow the settings provided in [13] and use $t_D = 1$ h and $r = 100$ m, the graph is built by setting $\delta = 200$ m. The number of past time slots is selected to be $\kappa = 72$ hours and the number of time slots to forecast is set to $\tau = 24$ hours. After processing, the description of the datasets is presented in Table I.

### B. Experimental Setting

To evaluate the proposed method, we divide the known entries into train, validation and test sets. To reflect real-time conditions where only past data is known at the time of forecasting, the validation and test data points are selected to be at a later date than the training data points. Specifically for the CoT dataset, the training set is composed of the data points collected before May 21, 2018, the validation set contains data points collected between May 22-23, 2018, and the testing set has data between May 24-31, 2018. Similarly, for the Kunak dataset, the test set contains data between April 24-30, 2021, the validation set has data between April 17-23, 2021, and the rest is used for training. We compute three common evaluation metrics, namely, the mean absolute error (MAE), the root mean squared error (RMSE) and the correlation coefficient.

Certain parameters of the model are selected experimentally through a random search optimization process: we find the best results are obtained with one $f_{GCN}$ layer on prior, recognition and generation networks, $100$ epochs and an initial learning rate of $0.001$ and $0.0001$ for CoT and Kunak datasets, respectively. We make use of early stopping to avoid overfitting and a batch size of 12. The tuning parameters of the loss function are set experimentally: the coefficient for correlation loss to $\alpha = 1$, the KL divergence coefficient to $\beta = 0.5$, the temporal smoothness coefficient to $\gamma = 0.001$, the weight decay to $\lambda = 0.00001$ and the temporal neighbourhood width to $w_T = 3$. The dropout rate is set to $0.5$ on the GCN and GRU layers in the encoder and decoder. In the computation of the matrix of context embeddings $\boldsymbol{C}$, we set the dimensionality of all the embeddings (POIs, weather, timestamp, past pollution and context) to $d = 100$. For all GCN layers, we use a dimensionality of $D = 512$. We employ the ReLU activation function to activate the linear and GCN layers, except for the last GCN layer of the $\boldsymbol{\sigma}$ branches where the $\tanh(\cdot)$ is used since data is normalised in this range.

We test our model's performance with respect to the models presented in Section II. First, we compare against two baseline models, namely, the GRU+GCN [17] deep forecasting model and the statistical ARIMA [2]. Additionally, we compare against state-of-the-art VAE models such as the original VAE [18], VGAE [23] and conditional VAE [25]. All models are trained in the same training set as the proposed model.

### C. Results and Analysis

The forecasting performances are reported in Table II for CoT and Kunak datasets. Note that the generative step of VAE-based models is inherently randomized, i.e., sampling $\boldsymbol{z}$ twice from $q(\boldsymbol{z}|\boldsymbol{c})$ will not yield the same results. For robustness, we employ 4-fold cross validation on the VAE-based models and report the standard deviation of these results. ARIMA [2] provides the worst forecasting accuracy; this is because this model follows a statistical approach where domain knowledge and context data are not used. In contrast to statistical models, [17] assumes the existence of latent embeddings which characterize discrete locations, leading to better performance results. It is evident that, in general,

TABLE II
AIR QUALITY $NO_2$ FORECASTING RESULTS.

| | CoT | | | Kunak | | |
|---|---|---|---|---|---|---|
| | MAE | RMSE | Corr Coef | MAE | RMSE | Corr Coef |
| Imputation+ARIMA [2] | $23.961 \pm 0.069$ | $17.123 \pm 0.041$ | $0.214 \pm 0.011$ | $23.549 \pm 1.026$ | $28.631 \pm 1.142$ | $-0.266 \pm 0.131$ |
| GRU+DNN [17] | $20.481 \pm 0.085$ | $25.049 \pm 0.088$ | $0.303 \pm 0.0012$ | $17.863 \pm 0.971$ | $23.010 \pm 1.057$ | $0.054 \pm 0.124$ |
| VAE [18] | $17.304 \pm 0.083$ | $21.696 \pm 0.124$ | $0.257 \pm 0.001$ | $16.915 \pm 0.631$ | $22.069 \pm 0.051$ | $\mathbf{0.275 \pm 0.038}$ |
| VGAE [23] | $16.810 \pm 0.036$ | $21.199 \pm 0.042$ | $0.275 \pm 0.006$ | $17.355 \pm 0.731$ | $22.587 \pm 0.631$ | $0.220 \pm 0.038$ |
| CVAE [26] | $17.189 \pm 0.058$ | $22.509 \pm 0.071$ | $0.346 \pm 0.004$ | $17.355 \pm 0.531$ | $22.587 \pm 0.606$ | $0.099 \pm 0.059$ |
| cVGAE* | $\mathbf{15.899 \pm 0.079}$ | $\mathbf{20.319 \pm 0.051}$ | $\mathbf{0.361 \pm 0.049}$ | $16.534 \pm 0.019$ | $21.530 \pm 0.004$ | $0.209 \pm 0.046$ |
| cVGAE Autoregressive* | $20.769 \pm 0.089$ | $26.230 \pm 0.119$ | $0.305 \pm 0.006$ | $\mathbf{15.553 \pm 0.135}$ | $\mathbf{20.671 \pm 0.262}$ | $0.247 \pm 0.099$ |

VAEs obtain better forecasting accuracy; we argue this is due to their encoder-decoder scheme. Amongst them, our cVGAE models—denoted by $^\star$ in Table II—achieve the best performance for RMSE and MAE while reaching comparable results in the correlation coefficient score. While the cVGAE autoregressive model outperforms other methods in the Kunak dataset, it shows worse performance in the CoT dataset. We argue this is due to the small size of the CoT dataset, which results in the model overfitting in datasets of short temporal spans.

## V. CONCLUSION

Mobile stations are a promising approach to instantly measure air pollutant concentrations. The collected measurements have high spatial density but suffer from low temporal resolution at each location. We argue that learning the latent distribution of past and context data per location with a VAE architecture allows to solve the issue. In this setting, we formulated the air quality forecasting problem as a graph-based matrix completion problem. To solve it, a variational graph autoencoder with condition was proposed. The presented model has shown to deliver high-quality forecasting by effectively capturing the context knowledge and the spatio-temporal correlations in the measurements. The model outperforms state-of-the-art methods in air quality forecasting. Future work will aim at generalizing the application domain of the cVGAE to other graph-structured real-world data (e.g., social media data, IoT data).

## REFERENCES


[1] J. Chen *et al.*, "Seasonal modeling of PM2.5 in California's San Joaquin Valley," *Atmospheric Environment*, vol. 92, pp. 182–190, 2014.

[2] E. R. Ziegel, *Technometrics*, vol. 37, no. 2, pp. 238–239, 1995.

[3] Z. Ma *et al.*, "Estimating Ground-Level PM2.5 in China Using Satellite Remote Sensing," *Environmental Science & Technology*, vol. 48, no. 13, pp. 7436–7444, 2014.

[4] M. W. Gardner and S. R. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric Environment*, vol. 32, pp. 2627–2636, 1998.

[5] V. R. Prybutok, J. Yi, and D. Mitchell, "Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations," *Eur. J. Oper. Res.*, vol. 122, pp. 31–40, 2000.

[6] P. Perez and J. Reyes, "An integrated neural network model for PM10 forecasting," *Atmospheric Environment*, vol. 40, pp. 2845–2851, 2006.

[7] V. O. K. Li *et al.*, "Deep Learning Model to Estimate Air Pollution Using M-BP to Fill in Missing Proxy Urban Data," in *IEEE Global Communications Conference*, 2017, pp. 1–6.

[8] Z. Qi *et al.*, "Deep air learning: Interpolation, prediction, and feature analysis of fine-grained air quality," *IEEE Trans. on Knowl. and Data Eng.*, vol. 30, no. 12, p. 2285–2297, 2018.

[9] Y. Lin *et al.*, "Exploiting spatiotemporal patterns for accurate air quality forecasting using deep learning," in *Proceedings of the 26th ACM International Conference on Advances in Geographic Information Systems*, 2018, p. 359–368.

[10] L. Qi *et al.*, "Deep learning for air quality forecasts: a review," *Current Pollution Reports*, vol. 6, pp. 1–11, 2020.

[11] A. Baklanov and Y. Zhang, "Advances in air quality modeling and forecasting," *Global Transitions*, vol. 2, pp. 261–270, 2020.

[12] C.-J. Huang and P.-H. Kuo, "A Deep CNN-LSTM Model for Particulate Matter PM2.5 Forecasting in Smart Cities," *Sensors*, vol. 18, no. 7, 2018.

[13] T. H. Do *et al.*, "Graph-deep-learning-based inference of fine-grained air quality from mobile iot sensors," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8943–8955, 2020.

[14] D. Hasenfratz *et al.*, "Pushing the spatio-temporal resolution limit of urban air pollution maps," in *2014 IEEE International Conference on Pervasive Computing and Communications*, 2014, pp. 69–77.

[15] T. Coenen *et al.*, "City of things: An integrated and multi-technology testbed for IoT smart city experiments," in *IEEE International Conference on Smart Cities*, 2016, pp. 1–8.

[16] J. Hofman *et al.*, "Spatiotemporal air quality inference of low-cost sensor data: Evidence from multiple sensor testbeds," *Environmental Modelling and Software*, vol. 149, p. 105306, 2022.

[17] Y. Seo *et al.*, "Structured sequence modeling with graph convolutional recurrent networks," in *25th International Conference Neural Information Processing*, vol. 11301. Springer, 2018, pp. 362–373.

[18] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2014.

[19] K. K. Rani Samal *et al.*, "Data Driven Multivariate Air Quality Forecasting using Dynamic Fine Tuning Autoencoder Layer," in *IEEE 17th India Council International Conference*, 2020, pp. 1–6.

[20] A. Dairi *et al.*, "Integrated Multiple Directed Attention-Based Deep Learning for Improved Air Pollution Forecasting," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–15, 2021.

[21] X. Li *et al.*, "Deep learning architecture for air quality predictions," *Environmental Science and Pollution Research*, vol. 23, pp. 22 408–22 417, 2016.

[22] D. Liang *et al.*, "Variational autoencoders for collaborative filtering," in *Proceedings of the 2018 World Wide Web Conference*, 2018, p. 689–698.

[23] T. N. Kipf and M. Welling, "Variational graph auto-encoders," 2016.

[24] A. Hasanzadeh *et al.*, "Semi-implicit graph variational auto-encoders," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[25] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[26] A. Fanfarillo *et al.*, "Probabilistic forecasting using deep generative models," *GeoInformatica*, vol. 25, 2021.

[27] D. Rigoni, N. Navarin, and A. Sperduti, "Conditional constrained graph variational autoencoders for molecule design," 2020.

[28] Q. Liu *et al.*, "Constrained graph variational autoencoders for molecule design," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, p. 7806–7815.

[29] M.-H. Yu *et al.*, "Content learning with structure-aware writing: A graph-infused dual conditional variational autoencoder for automatic storytelling," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 7, pp. 6021–6029, 2021.