MATRIX COMPLETION WITH VARIATIONAL GRAPH AUTOENCODERS APPLICATION IN HYPERLOCAL AIR QUALITY INFERENCE

Tien Huu Do ✦ ☆, Duc Minh Nguyen ✦ ☆, Evaggelia Tsiligianni ✦ ☆, Angel Lopez Aguirre † ‡, Valerio Panzica La Manna ◊,

Frank Pasveer ◊, Wilfried Philips † ☆, Nikos Deligiannis ✦ ☆

✦Vrije Universiteit Brussel - †IPI Ghent University - ‡Escuela Superior Politecnica del Litoral – ☆imec Leuven – ◊Holst Center, imec

1. Problem Statement

Problem: Data from a mobile air pollution monitoring system are characterized by:

- High spatial resolution: the mobile stations can cover a larger geographical region compared to fixed stations.
- Low temporal resolution per measured location: the sensors are moving.

Goal: Air pollution inference at a high spatio-temporal resolution using data from a limited number of mobile and fixed stations.

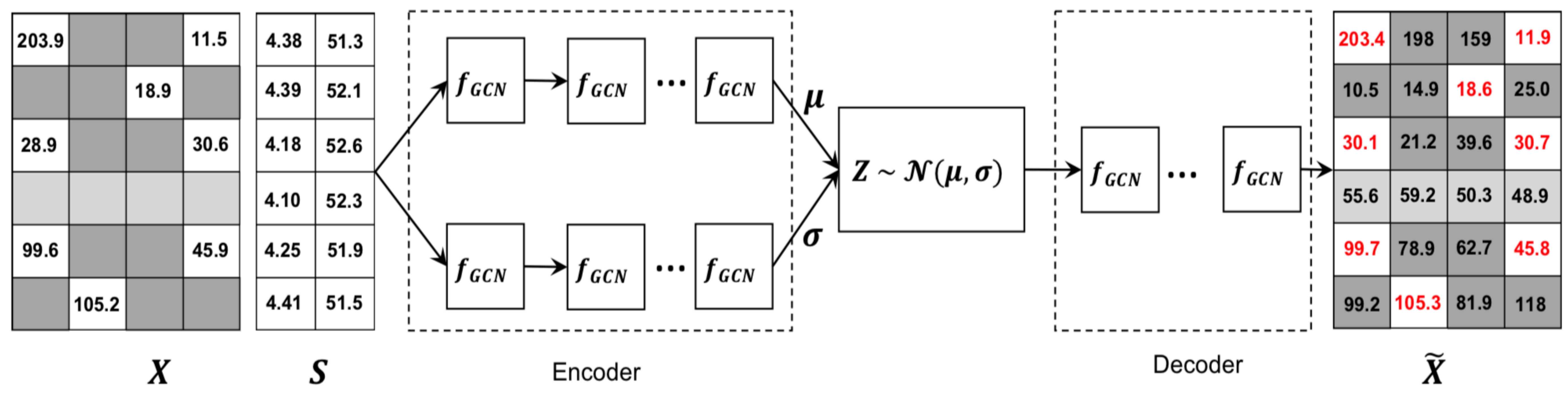
2. Proposed Method

Data preprocessing

- A number of locations of interest are defined in the road network of Antwerp.
- Measurements collected during a time interval τ, at a distance r of the considered locations are aggregated.
- Aggregated measurements from N locations at T timeslots are arranged in a matrix $X \in R^{N\times T}$. Only a few entries of X are known.

Formulation of matrix completion on graphs

- The locations of interest are considered as the nodes V of a graph $G=(V,E)$.
- Two nodes are connected if they belong to the same road segment or their distance is smaller than δ.
- The weight of a connection is the inverse of the geodesic distance between two locations.
- We utilize the graph to complete X.



Variational Graph Autoencoder for Air Quality Inference (AVGAE)

We propose a probabilistic autoencoder that incorporates graph information to learn the distribution of air pollution data.

$$\boldsymbol{\mu} = \mathrm{GCN}_{\mu}(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\Theta_1})$$
$$\boldsymbol{\sigma} = \mathrm{GCN}_{\sigma}(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{\Theta_2})$$
$$\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$$
$$\tilde{\boldsymbol{X}} = \mathrm{GCN}_{z}(\boldsymbol{Z}, \boldsymbol{\Phi})$$

The graph convolutional layer [1, 2]

$$f_{\mathrm{GCN}}(\boldsymbol{X}) = \sigma(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{W})$$
$$\tilde{A} = A + I_N$$
$$\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$$

A: The adjacency matrix

W: The weight matrix

The proposed loss function

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\Phi}) = \frac{1}{|\Omega|}\sum_{(i,j)\in\Omega}|\tilde{\boldsymbol{X}}_{ij} - \boldsymbol{X}_{ij}| +$$
$$\beta\mathcal{D}[q(\boldsymbol{Z}|\boldsymbol{X})\|p(\boldsymbol{Z})] + \gamma\sum_{(i,j)}\sum_{k\in\mathcal{T}(i,j)} e^{-|j-k|}(\tilde{\boldsymbol{X}}_{ij} - \tilde{\boldsymbol{X}}_{i,k})^2$$

3. Experimental Results

Table 1: The description of NO2 and PM2.5 datasets.

	NO_2	$PM_{2.5}$
Number of locations	3630	4086
Duration in hour	720	720
Max concentration	633.65	189.03
Min concentration	0.16	0.07
Mean concentration	85.50	9.83
% of known entries versus all	0.60	0.56

Table 2: Air quality inference result

	NO_2		$PM_{2.5}$	
	MAE	RMSE	MAE	RMSE
Kriging linear	18.9	28.43	3.28	7.98
Kriging exponential	15.86	25.58	2.89	7.43
KNN collaborative filtering	20.92	32.67	3.60	7.47
SVD	27.35	38.32	7.41	13.40
NMF	71.67	82.34	6.75	13.09
NMC	22.12	32.83	3.99	8.35
RGCNN	48.6	60.11	6.2	15.4
AVGAE (Proposed)	**14.92**	**24.33**	**2.56**	**6.42**

1) T. N. Kipf and M. Welling, "Variational graph auto-encoders," NIPS Workshop on Bayesian Deep Learning, 2016.
2) T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in ICML, 2017.
3) D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv:1312.6114, 2013.