AI4VN - 2022

Air Quality Forecasting Challenge

Team: HBLH

Members:

- Hoàng
- Bá
- Lộc
- Huỳnh

VNU UET - Institute for Artificial Intelligence

Mục lục

Giới thiệu

PM2.5

Chỉ số đo lường chất lượng không khí.

Bài toán

Chủ đề và bài toán của cuộc thi.

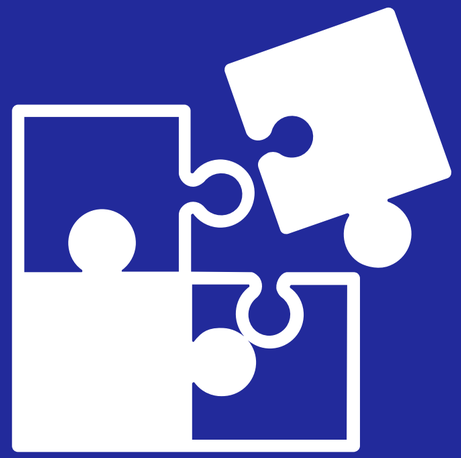
Dữ liệu

Thông tin bộ dữ liệu private.

PM2.5

- Chỉ số PM2.5 là **chỉ số quan trọng được sử dụng để đánh giá chất lượng không khí.**
- Nồng độ càng cao thì chất lượng không khí càng suy giảm, gây ảnh hưởng tới đường hô hấp và sức khỏe con người.

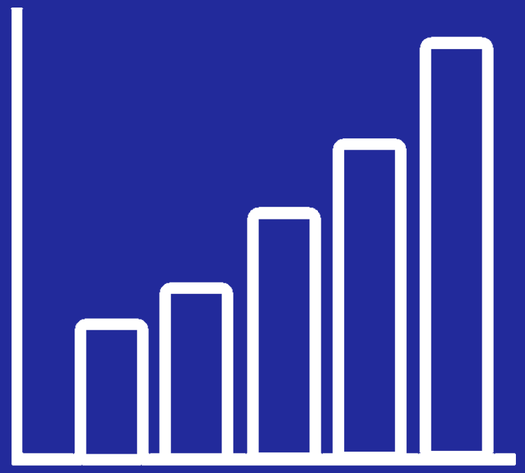
Bài toán



Khó khăn: Do chi phí và địa hình, nhiều khu vực không thể đặt các thiết bị để đo PM2.5.

Bài toán: Đưa ra được dự đoán về chỉ số PM2.5 tại 1 địa điểm bất kỳ được cung cấp tọa độ và có khả năng đưa ra dự báo về PM2.5 trong 24h trong tương lai.

Dữ liệu



- Dữ liệu Training
  - Dữ liệu chất lượng không khí (**71 trạm x 6000 timesteps x 3 features**).
  - Dữ liệu khí tượng (**143 trạm x 2000 timestep x 5 features**).
- Dữ liệu Input Test (89 folders):
  - Dữ liệu chất lượng không khí (**8 trạm x 168 timesteps x 3 features**).
  - Dữ liệu khí tượng (**143 trạm x 56 timesteps x 5 features**).

Giải pháp

Data Preprocessing

Quá trình tiền xử lý dữ liệu.

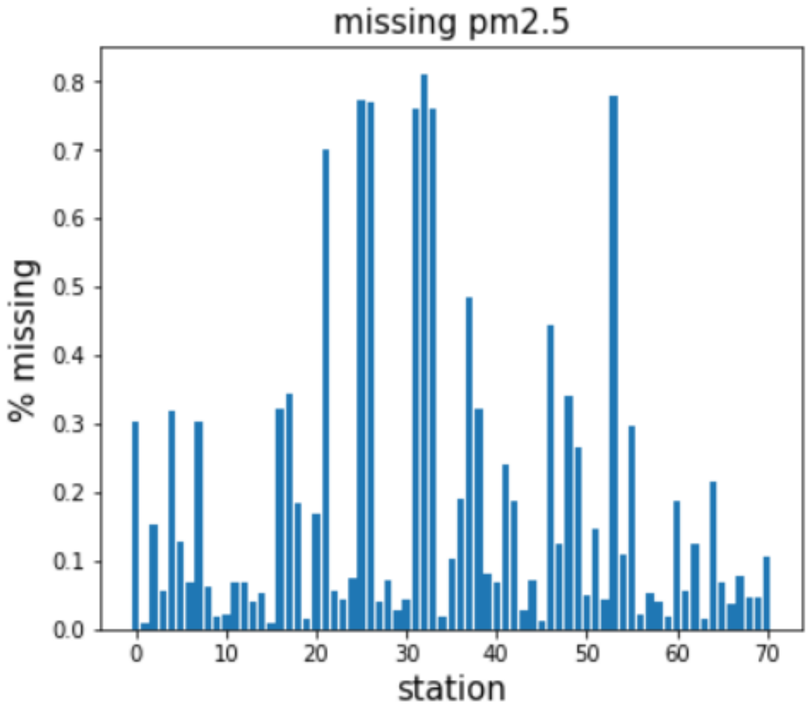
Methedology

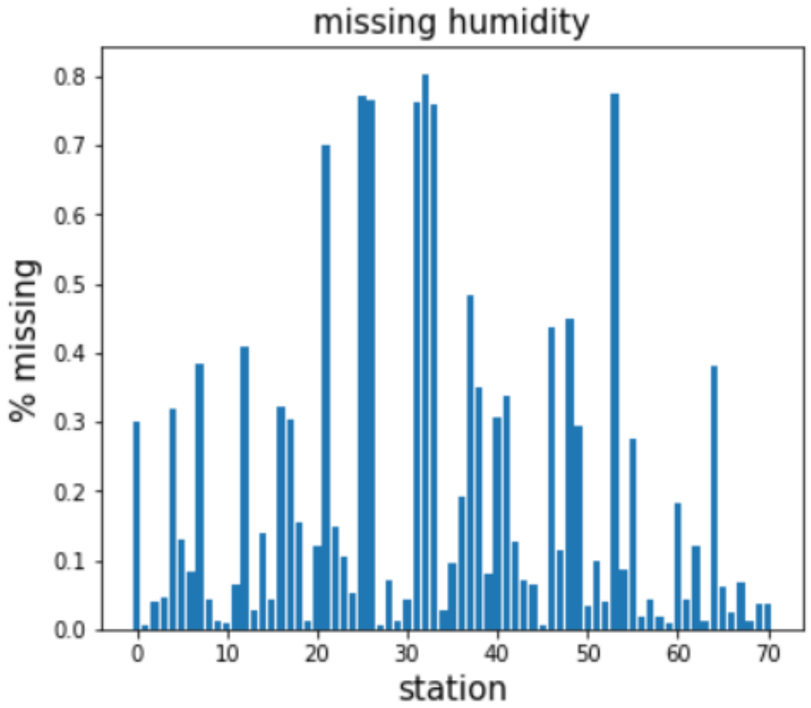
Đi sâu vào model và
những phương pháp được sử dụng.

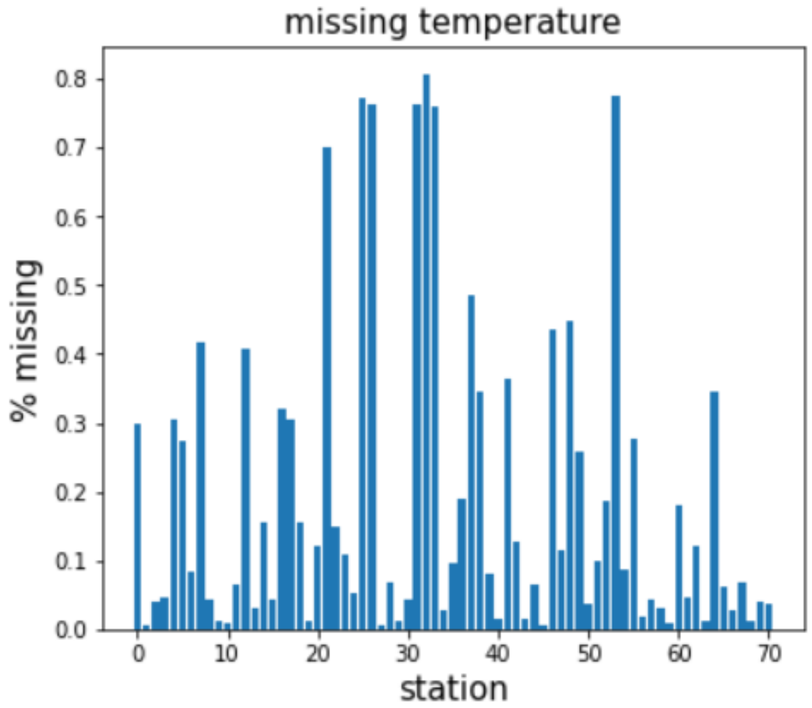
Data Preprocessing

1. Data Understanding
2. Data Imputation
3. Convert Meteorology Data
4. Data Normalization

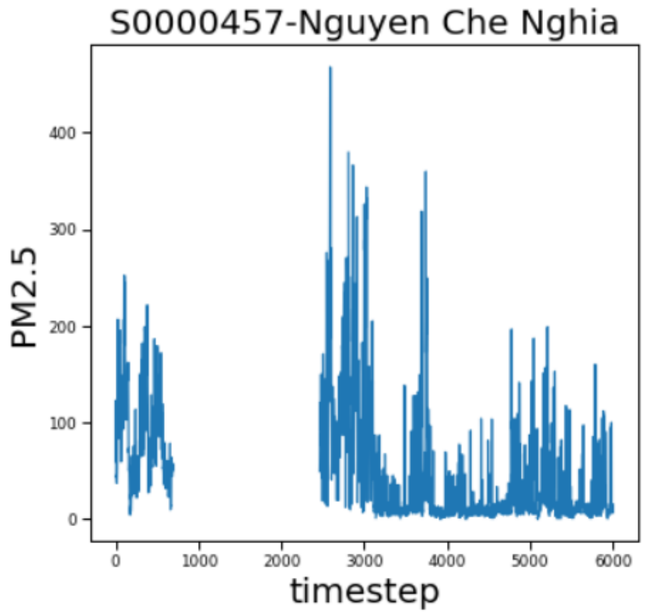
Data Understanding

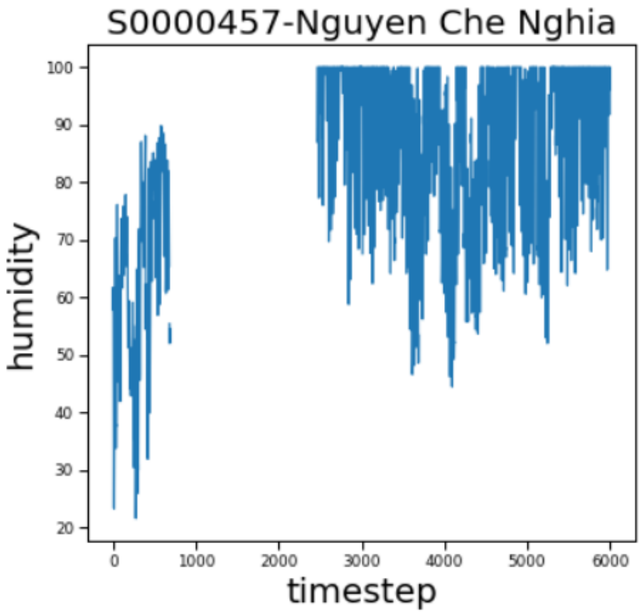


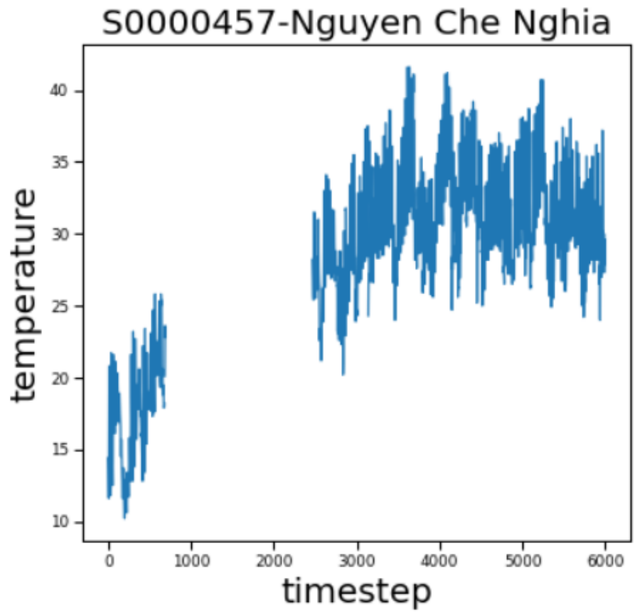




Data Understanding







Data Imputation

- **Drop Missing Data**

  Bỏ những dữ liệu bị mất.

- **SpLine Interpolate**

  Phương pháp điền nội suy.

- **Median**

  Điền trung vị.

- **Inversed Distance Weighted - IDW**

  Sử dụng khoảng cách giữa các trạm.

$$AQI_i = \frac{\sum_j^n AQI_j \times \text{distance}^{-\beta}(i,j)}{\sum_j^n \text{distance}^{-\beta}(i,j)}$$

Thí nghiệm

So sánh hiệu quả của các phương pháp fill missing data.

Metric	IDW	SpLine	Median
MAE	42.37	39.2	49.9
RMSE	61.63	58.87	65.39

Convert Meteorology Data

- **Từ hai vector u và v biển đổi ngược lại sang 2 giá trị:**
  - Wind Speed
  - Wind Direction

- **Sử dụng công thức:**
  - Wind speed:

$$ws = \sqrt{u^2 + v^2}$$

  - Wind direction:

$$\text{atan2}(y, x) = \begin{cases} \arctan\frac{y}{x} & x > 0 \\ \arctan\frac{y}{x} + \pi & y \geq 0, x < 0 \\ \arctan\frac{y}{x} - \pi & y < 0, x < 0 \\ +\frac{\pi}{2} & y > 0, x = 0 \\ -\frac{\pi}{2} & y < 0, x = 0 \\ \text{undefined} & y = 0, x = 0 \end{cases}$$

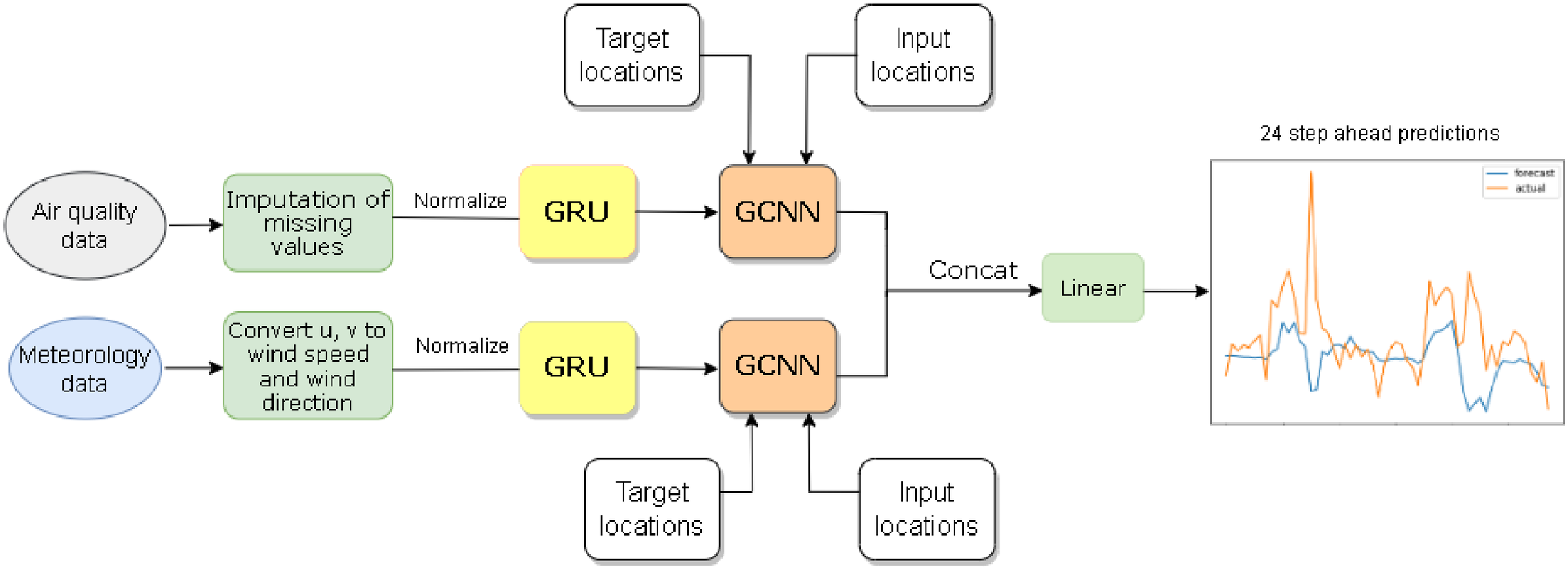
Data Normalization

- Sử dụng Standard Normalization cho tập dữ liệu.

$$z = \frac{x_i - \mu}{\sigma}$$

Methedology

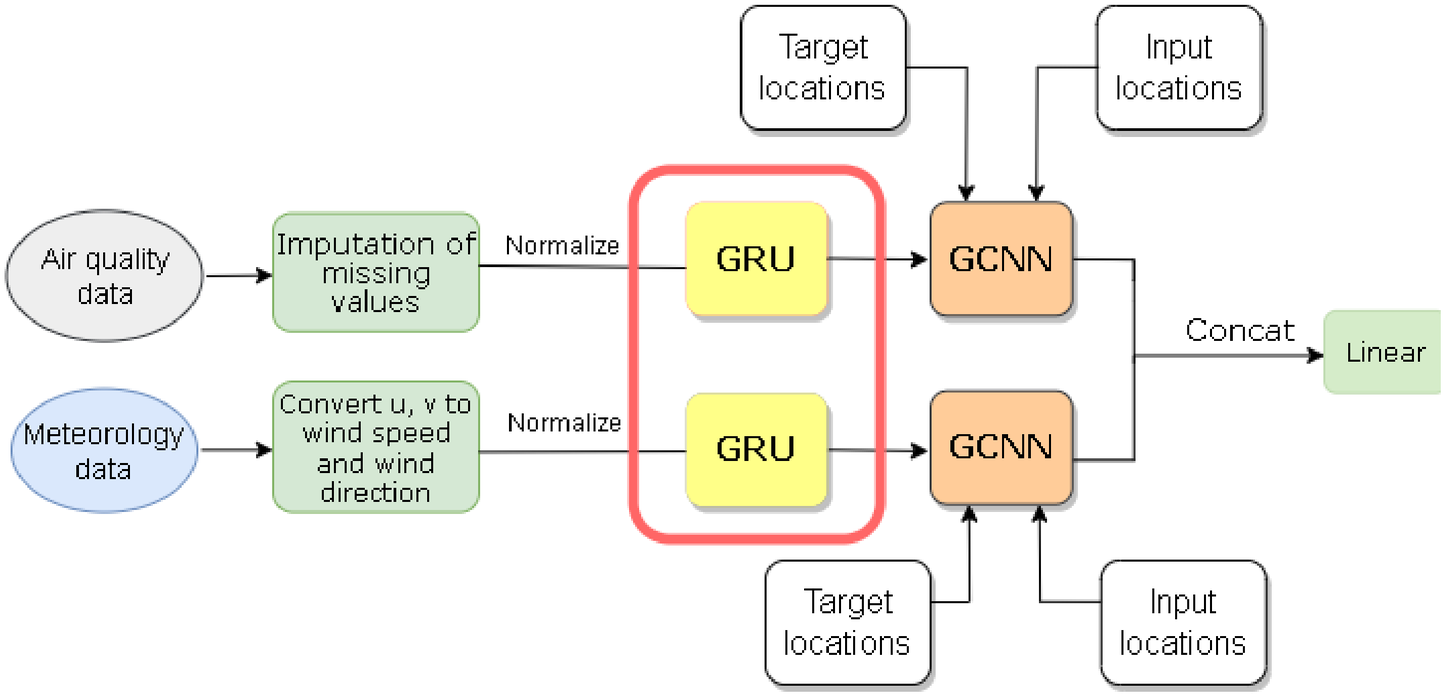
Sơ đồ Mô hình



BiGRU (Bidirectional Gated Recurrent Unit)

- Feature extractor của dữ liệu dạng timeseries
- Vì timesteps của tập **dữ liệu khí tượng** và **dữ liệu không khí** bị chênh lệch nhau nên sử dụng 2 GRU để extract cho từng tập dữ liệu tương ứng.

- **Input:**
  - **Data processed (station, sequence, feature)**
- **Output:**
  - **(station, feature_extracted)**



Kết quả thí nghiệm

	Bidirectional	Undirectional
MDAPE	**0.73**	0.76
MAE	**28.47**	28.98
MAPE	**1.09**	1.15
R2	**-0.41**	-0.49
RMSE	**40.72**	41.66

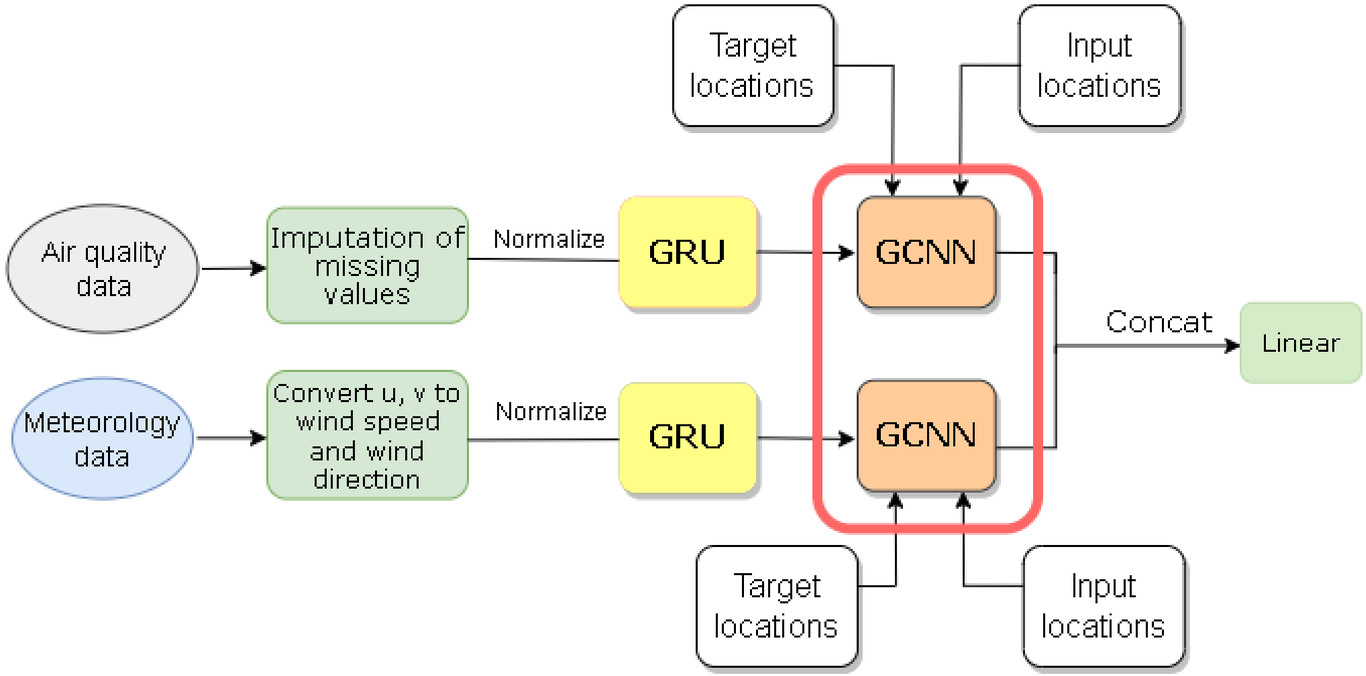
	GRU	LSTM
MDAPE	**0.73**	1.05
MAE	**28.47**	30.5
MAPE	**1.09**	1.78
R2	**-0.41**	-0.52
RMSE	**40.72**	41.82

Graph Conv

- **Input:**
  - **Input locations (input_station, [longitude, latitude])**
  - **Target locations (target_station, [longitude, latitude])**
  - **Output GRU (input_station, feature_extractor)**
- **Output:**
  - **(target_station, feature_extractor)**

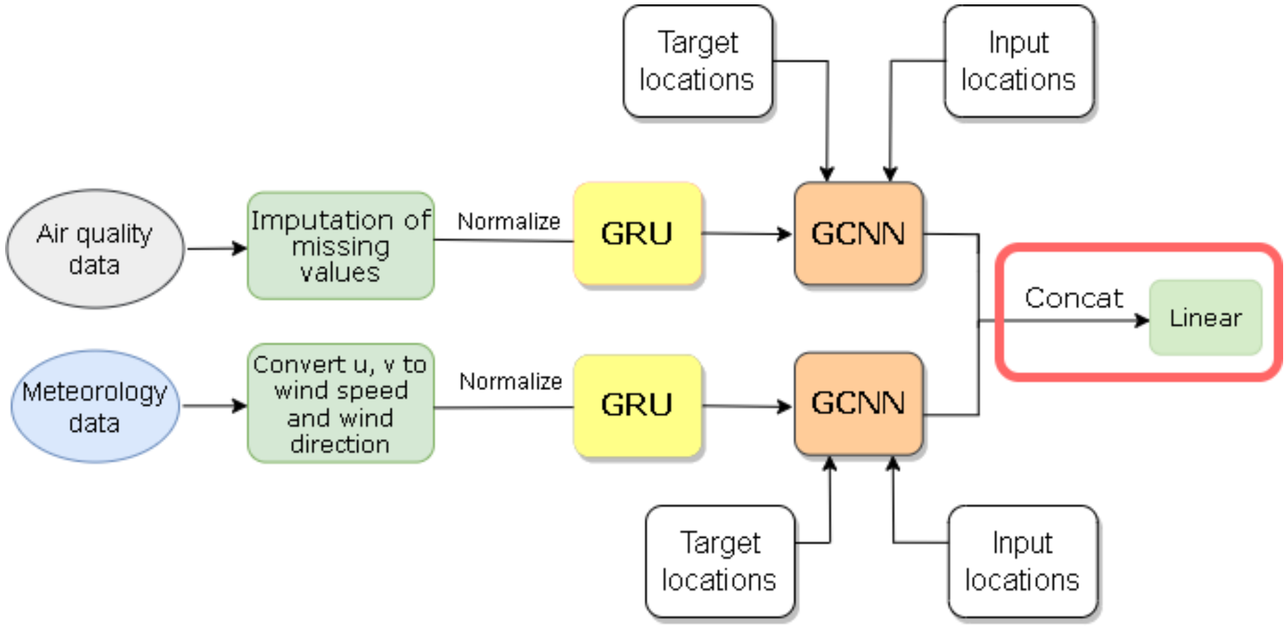
$$\mathbf{x}_i' = \mathbf{W}_1 \mathbf{x}_i + \mathbf{W}_2 \sum_{j \in \mathcal{N}(i)} e_{j,i} \cdot \mathbf{x}_j$$

- Tổng hợp feature từ các trạm nguồn cho các trạm đích.



Concat + Linear

- Input:
  - Output GCNN (Target station, feature_extractor)
- Output:
  - (Target station, 24)
- Concat: Tổng hợp feature từ 2 nguồn khí tượng và chất lượng không khí
- Linear: Dự đoán cho 24h tiếp theo



Setup Thí nghiệm

Training set

Setup dữ liệu training cho mô hình.

Metrics

Những phương pháp đánh giá sử dụng.

Training set

- Có 71 Trạm đo, ta chia ra:
  - 10 Trạm để đánh giá
  - 61 Trạm được luân phiên chia ra để dùng làm input và target
- Trong đó, 51 trạm input - 10 trạm target random theo từng batch.

Kết quả thí nghiệm

	Chia random	Cố định
MDAPE	**0.73**	0.72
MAE	**28.47**	30.6
MAPE	**1.09**	1.03
R2	**-0.41**	-0.72
RMSE	**40.72**	43.11

Metrics

MDAPE

$$MDAPE = median\left(\left|\frac{y_t - \hat{y}_t}{y_t}\right|\right)$$

R2

$$R2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y}_t)^2}$$

MAPE

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|$$

MAE

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$$

RMSE

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

Cải tiến

The First Law of Geography (Tobler 1970):

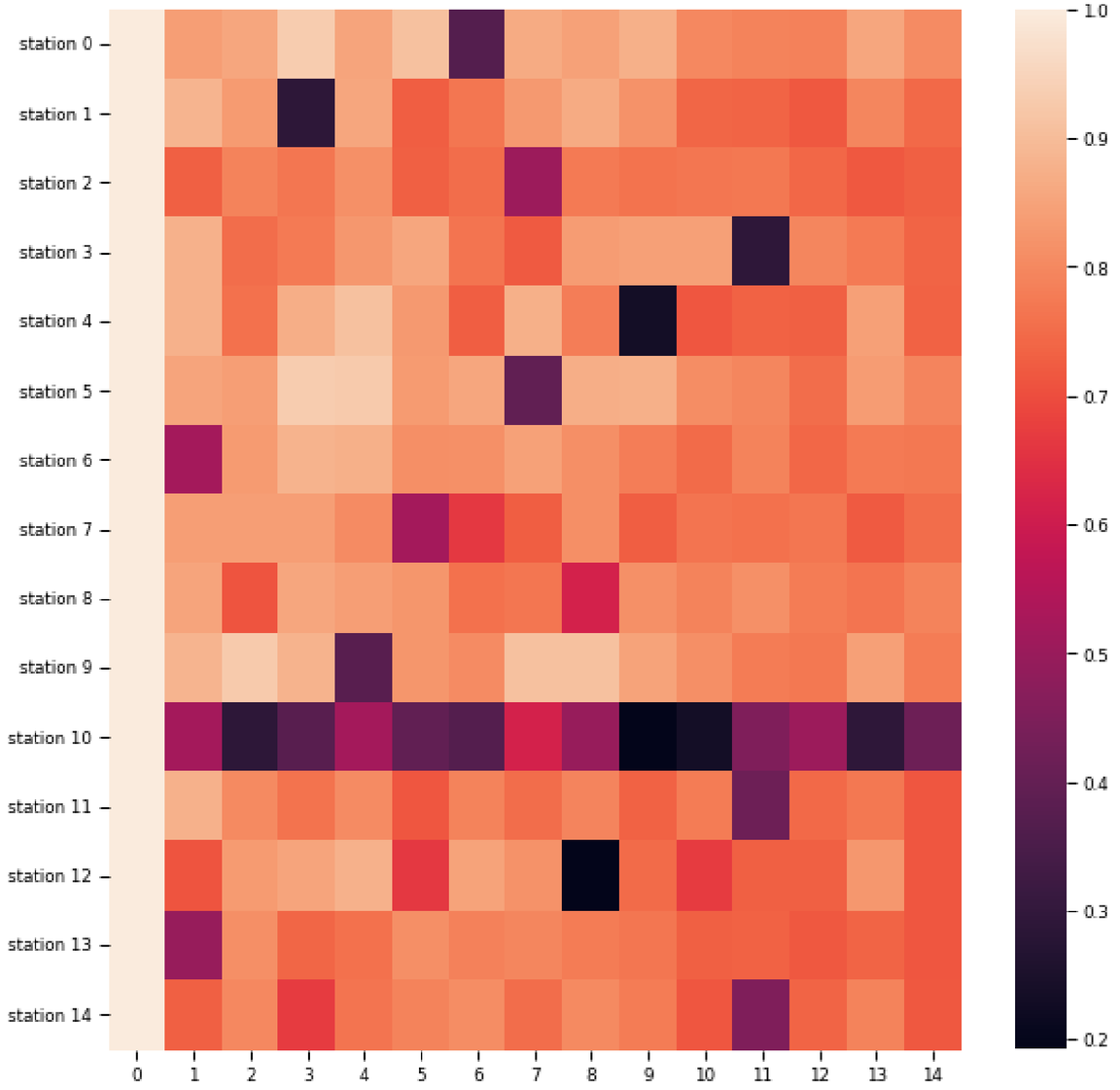
‘Everything is related to everything else, but near things are more related than distant things’

The Third Law of Geography (Zhu 2018):

‘The more similar geographic configurations of two points (areas), the more similar the values (processes) of the target variable at these two points (areas)’.

Biểu đồ tương quan giữa nồng độ PM2.5 của 14 trạm trong tập dữ liệu chất lượng không khí.

Kết luận: *Sự tương quan giữa các trạm không chỉ phụ thuộc vào khoảng cách mà còn phụ thuộc vào tính chất của môi trường xung quanh trạm đó.*



Kết luận

01

Thí nghiệm cho thấy **IDW** và **SpLine** là hai phương pháp fill missing data hiệu quả nhất.

02

Áp dụng **BiGRU** cho **Feature Extractor** (đối với data nhỏ và có sequence vừa).

03

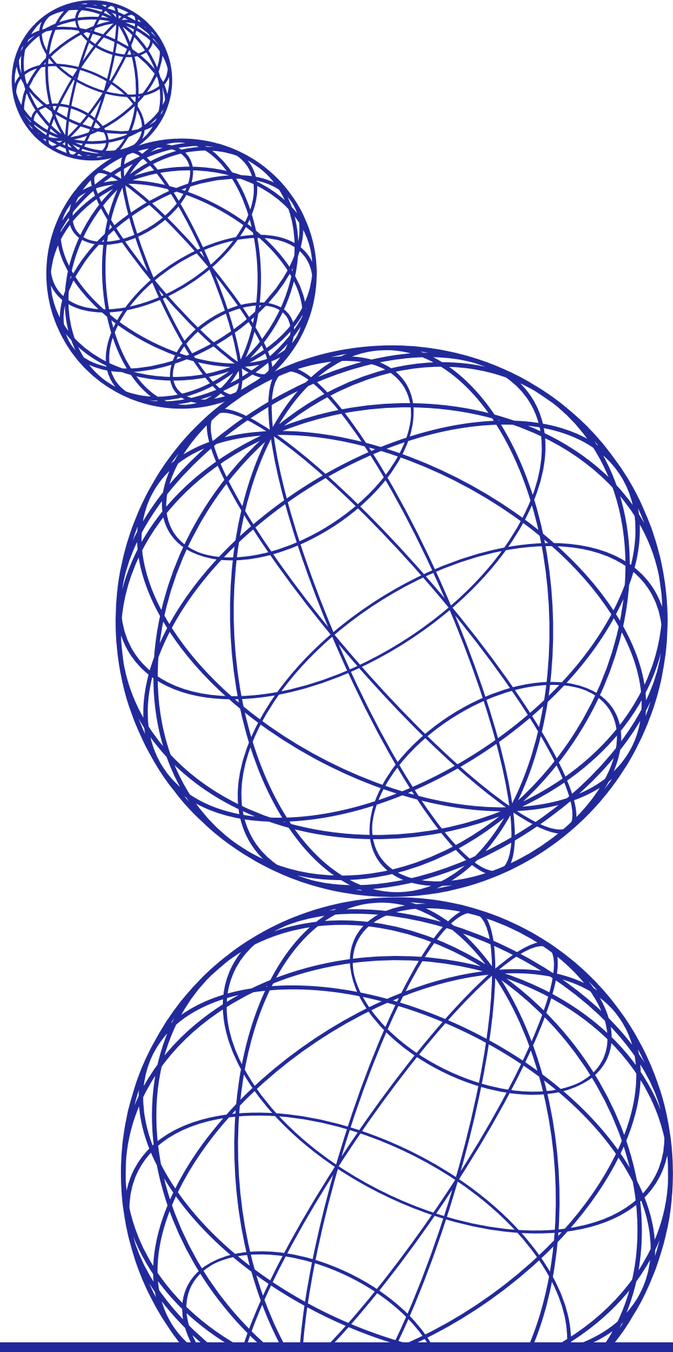
Tổng hợp các features cho trạm nguồn bằng cách sử dụng **Graph Conv (GCNN)**.

04

Khi training, nếu chia dữ liệu các trạm **một cách random** sẽ hiệu quả hơn để cố định.

05

Có thể cải tiến mô hình bằng cách thêm **dữ liệu về môi trường** (*land use, road network, . . .*)



Thank you!

Do you have any questions?