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 \( \tau \), 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 \( \delta \).
- 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.

\[
\begin{align*}
\mu &= \text{GCN}_\mu(X, S, \Theta_1) \\
\sigma &= \text{GCN}_\sigma(X, S, \Theta_2) \\
Z &= \mathcal{N}(\mu, \sigma) \\
\tilde{X} &= \text{GCN}_Z(Z, \Phi)
\end{align*}
\]

The graph convolutional layer \([1, 2]\)  
\[
f_{\text{GCN}}(X) = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} X W) \\
\tilde{A} = A + I_N \\
\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}
\]

\( A \): The adjacency matrix  
\( W \): The weight matrix

The proposed loss function

\[
\mathcal{L}(X, \Theta_1, \Theta_2, \Phi) = \frac{1}{|\Omega|} \sum_{(i,j) \in \Omega} |\tilde{X}_{ij} - X_{ij}| + \\
\beta D[q(Z|X)||p(Z)] + \gamma \sum_{(i,j) \in \Omega} \sum_{k \in \mathcal{T}(i,j)} e^{-|j-k|} |\tilde{X}_{ij} - \tilde{X}_{ik}|^2
\]

3. Experimental Results

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

<table>
<thead>
<tr>
<th></th>
<th>NO2</th>
<th>PM2.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of locations</td>
<td>3630</td>
<td>4086</td>
</tr>
<tr>
<td>Duration in hour</td>
<td>720</td>
<td>720</td>
</tr>
<tr>
<td>Max concentration</td>
<td>633.65</td>
<td>189.03</td>
</tr>
<tr>
<td>Min concentration</td>
<td>0.16</td>
<td>0.07</td>
</tr>
<tr>
<td>Mean concentration</td>
<td>85.50</td>
<td>9.83</td>
</tr>
<tr>
<td>% of known entries versus all</td>
<td>0.60</td>
<td>0.56</td>
</tr>
</tbody>
</table>

**Table 2:** Air quality inference result

<table>
<thead>
<tr>
<th></th>
<th>NO2</th>
<th>PM2.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mae</td>
<td>14.92</td>
<td>24.33</td>
</tr>
<tr>
<td>RMSE</td>
<td>22.12</td>
<td>60.11</td>
</tr>
<tr>
<td>KNN</td>
<td>22.12</td>
<td>60.11</td>
</tr>
<tr>
<td>SVD</td>
<td>27.35</td>
<td>38.32</td>
</tr>
<tr>
<td>NMF</td>
<td>71.67</td>
<td>82.34</td>
</tr>
<tr>
<td>NMC</td>
<td>48.6</td>
<td>4086</td>
</tr>
<tr>
<td>AVGAE (Proposed)</td>
<td>14.92</td>
<td>24.33</td>
</tr>
</tbody>
</table>