Estimating Ground-level Nitrogen Dioxide Concentration From Satellite Data

Authors:

Bibhash Pran Das¹, Muhammad Salman Pathan² †, Yee Hui Lee³, and Soumyabrata Dev²,4,5

¹National Institute of Technology Rourkela, India
²ADAPT SFI Research Centre, Dublin, Ireland
³School of Electrical and Electronic Engineering, Nanyang Technological University Singapore
⁴School of Computer Science, University College Dublin, Ireland.
⁵Beijing-Dublin International College, Beijing, China.
†Presenting author and corresponding author
OVERVIEW

- Introduction
- Motivation & Problem Formulation
- Methodology
- Results & Analysis
- Conclusions & Future Work
- References
• Nitrogen Dioxide (NO$_2$) is an atmospheric trace gas that plays a major role in atmospheric chemistry, air pollution and climate
• High concentration of NO$_2$ is toxic to humans and may cause decreased lung function and respiratory symptoms among other ailments.
• Monitoring of NO$_2$ concentrations is therefore crucial for air pollution control and air quality monitoring.
• Traditionally, in situ measurement strategies are used which require installation of dedicated equipment.
•Monitoring in remote locations becomes an issue and only point estimates can be performed.
• Satellite monitoring offers a truly global round the clock measurement alternative. However, measurement of vertical concentration profile is not possible.
• Therefore estimation methods to infer near-surface concentrations from vertical column density measurements are necessary
• Existing methods in the field mostly rely on atmospheric chemistry models like GEOS-Chem, RAMS-CMAQ etc. or regression models like linear regression or gradient boosting methods.
• The introduction of deep learning methods to this field is relatively recent.
• Simple feed forward deep neural networks as well as residual deep neural networks have been used to estimate near-surface concentration values from satellite measurements of tropospheric vertical column density.
• Most existing methods however, perform a multivariate estimation i.e auxiliary variables such as wind speed, population density, elevation etc. are considered.
• Looking at global NO₂ concentration patterns, we can reason that neighbouring areas in general tend to have similar concentrations owing to similarity in demographics and geographical attributes.
• Therefore, exploiting these local patterns may provide a method to efficiently estimate near-surface NO₂ concentrations
• Convolutional Neural Networks are a class of deep learning models that use learnable filters to exploit local pixel dependencies to extract hierarchal features.

• For the task at hand, we reason that if concentrations are arranged in a two dimensional matrix format based on geographical locations, CNNs may be used to learn the mapping from satellite measured tropospheric vertical column density values to near surface concentrations.

• Therefore, in our work we propose a CNN architecture to perform an univariate estimation of near-surface concentrations. The study is carried out over Ireland.
A) Data Processing:

- Tropospheric vertical column densities are obtained from the Tropospheric Ozone Monitoring Instrument (TROPOMI) aboard ESA’s Sentinel 5-p satellite.
- The corresponding near-surface concentration values are obtained from Environmental Protection Agency Ireland’s monitoring station.
- These data are then processed using a gridding procedure to obtain 2-D matrices that represent the daily average near surface concentrations and tropospheric VCD respectively.
- The resolution of the grid is fixed at 0.05° latitude x 0.05° longitude
- The data collected is for the period of January 2020 to May 2021 and is over an area above Ireland
- Geo-spatial bounding box with lower left corner at (−9.40°E, 51.83°N) and upper right corner at (−4.03°E, 54.32°N) is selected to mark the area of study.
A) Data Processing:

- Nearest satellite or ground data point from any grid cell in the matrix is determined using geographical distance. For this purpose the Haversine formula is used.

\[
d = 2\pi \arcsin \left( \sin \left( \frac{\phi_2 - \phi_1}{2} \right) \cos \left( \phi_1 \right) \cos \left( \phi_2 \right) \sin \left( \frac{\lambda_2 - \lambda_1}{2} \right) \right)
\]

- Once the algorithm is repeated for every cell in the grid, and all such grids obtained for a particular day are averaged, we obtain 2D matrices representing the daily average concentration measured in ground monitoring station and a grid of tropospheric VCD values measured from TROPOMI.
B) Proposed Architecture:

- The proposed architecture is a convolutional neural network with 4 convolutional layers, a pooling layer between the 2nd and 3rd Conv layers and 2 dense layers to perform the regression.
- The network takes matrices of size 49x67 as input and regresses an output vector of size 3283 which is then reshaped back to 49x67 during post processing.
- Thus we have tropospheric VCD matrices as input and estimated near surface concentration matrices as output.
C) Training:

- The data preparation step presents us with a dataset of 485 samples. Since we have both the inputs and targets, the problem is approached as a supervised learning problem.
- A 5 Fold cross validation strategy is used for model training and validation.
- The model is trained for 50 epochs with an Adam optimizer with learning rate of 0.001 and a Mean Absolute Error (MAE) loss function.
- Root Mean Squared Error (RMSE) is chosen as the validation metric since it gives a measure of the deviation of the predicted near surface concentration values from the actual values.
- Validation scores in each of the 5 folds are recorded and the average of the 5 validation scores so obtained is reported as the cross validation score.
A) Model Evaluation:

- The model is evaluated using multiple metrics namely, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Pearson Coefficient averaged over days ($R^2_{DAvg}$).

- We believe that rather than calculating a simple Pearson Correlation coefficient, it makes sense to calculate correlation between predicted and true values individually for each day, instead of together for every day in the period.

- The average RMSE score so obtained is $7.20 \mu g/m^3$. Similarly a MAE of $4.94 \mu g/m^3$ was obtained.

- A $R^2_{DAvg}$ of 0.6493 is obtained.
B) Performance Analysis:

- The proposed method is benchmarked against XGBoost, Light Gradient Boosting Machines (LGBM) and Linear regression models.

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE Score</th>
<th>MAE</th>
<th>$R^2_{D_{avg}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed CNN</td>
<td>7.20</td>
<td>4.94</td>
<td>0.649</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>8.595</td>
<td>6.271</td>
<td>0.0806</td>
</tr>
<tr>
<td>XGBoost</td>
<td>8.534</td>
<td>6.217</td>
<td>0.046</td>
</tr>
<tr>
<td>Light GBM</td>
<td>8.538</td>
<td>6.221</td>
<td>0.044</td>
</tr>
</tbody>
</table>

- As with any deep learning model, the proposed method is certainly effected by the amount of data. To observe this effect an experiment was done by iteratively sampling the dataset and recording the performance of the model in each iteration.
B) Performance Analysis:

- 10% of the total dataset is selected by means of random sampling.
- A model with the proposed architecture is trained and validated on the sampled data using a 3 Fold cross validation strategy. The average RMSE score is recorded.
- Steps (1) and (2) are repeated 20 times. In each repetition the randomly sampled dataset varies. Therefore a spread of RMSE score values are obtained.
- The list of values obtained in step (4) are plotted in a box-plot.
- Steps (1) through (4) are repeated by gradually varying the sampling size from 10% to 100%.
CONCLUSION & FUTURE WORK

• In this work we presented a convolutional neural network based method to perform the estimation of near-surface NO$_2$ concentration over an area over Ireland using Tropospheric Vertical Column Density values measured from TROPOMI.

• The method addressed the problem as an univariate regression task and only relied on the VCD values as regressors.

• The proposed method demonstrates very competitive RMSE scores even though no auxiliary variables like demographic parameters, climate parameters of geographic parameters are used.

• In the future the effect of including other parameters such as ground-based weather observations along with tropospheric VCD values may be explored.

• The method can also be implemented to consider a larger area of study. Moreover this method may also be extended to perform studies of other atmospheric constituents such as O$_3$, SO$_2$, PM$_{2.5}$ etc.
REFERENCES

Thank You