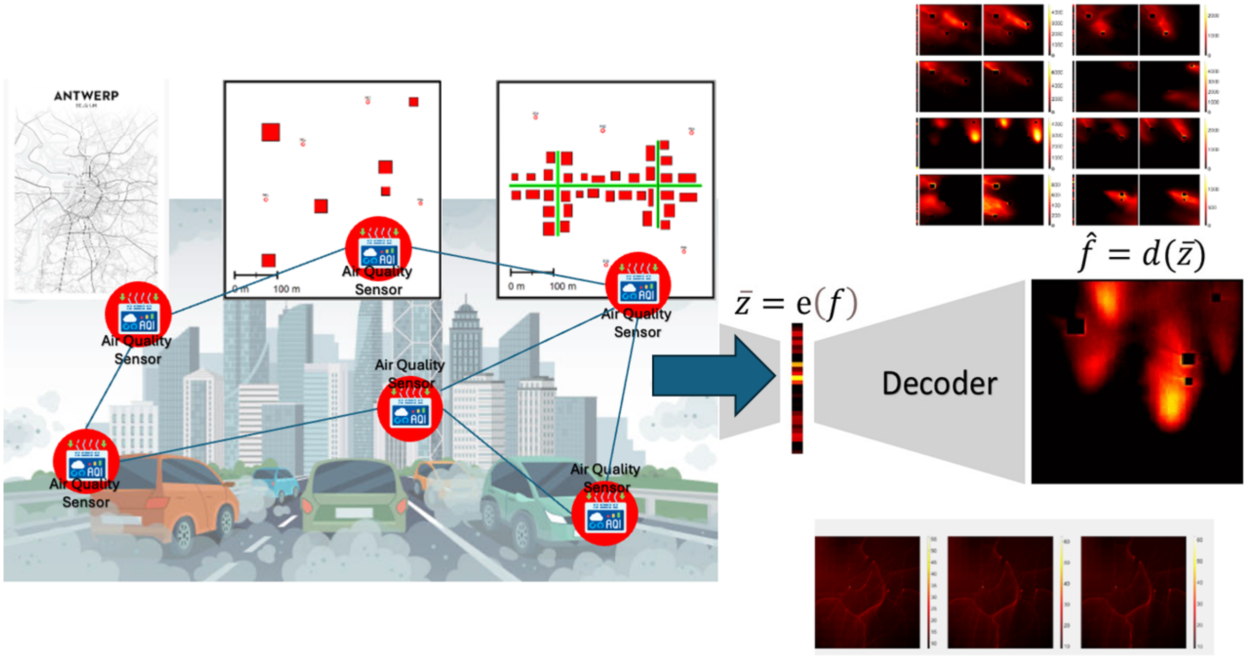
Graphical Abstract

Ridiculously Simple Data-Driven Air Pollution Interpolation Method

Alon Feldman, Shai Kendler, Enrico Pisoni, Barak Fishbain



Highlights

Ridiculously Simple Data-Driven Air Pollution Interpolation Method

Alon Feldman, Shai Kendler, Enrico Pisoni, Barak Fishbain

- The study presents an air pollution interpolation method using machine education.
- Hypothetical and real-world data from Antwerp, Belgium validate the approach.
- The method outperforms traditional techniques for creating dense pollution maps.

Ridiculously Simple Data-Driven Air Pollution Interpolation Method

Alon Feldman[a], Shai Kendler[b,c], Enrico Pisoni[d], Barak Fishbain[b,*]

[a] *Faculty of Mathematics, The Technion - Israel Institute of Technology, Rabin Hall, Technion City, Haifa, 320003, Israel*
[b] *Faculty of Civil and Environmental Engineering, The Technion - Israel Institute of Technology, Rabin Hall, Technion City, Haifa, 320003, Israel*
[c] *Environmental Physics Department, Israel Institute for Biological Research, P.O.B 19, Ness-Ziona, 7410001, Israel*
[d] *European Commission Joint Research Centre Ispra Sector, Via Enrico Fermi, 2749, Italy, Ispra, 21027, Italy*

Abstract

Air pollution interpolation is crucial for civil management: it is used to transform limited sensor data into comprehensive pollution maps. Various methods, including deterministic, geostatistical, and Machine Learning (ML)-based techniques have been utilized for this purpose. Deterministic methods rely on mathematical rules for estimation, whereas geostatistical techniques are based on spatial correlations. ML leverages historical data for predictions. Each method has limitations: deterministic methods do not adequately model environmental complexity, geostatistical methods struggle with small-scale areas, and ML depends heavily on data availability. On the other hand, air pollution simulators, which are driven by physicochemical dispersion models, capture intricate pollution dispersion patterns but are unsuitable for real-time interpolation. However, recent advancements in ML may offer potential solutions by integrating simulation data with ML methodologies to respond to interpolation needs.

This study introduces an air pollution interpolation approach combining simulated air-dispersion patterns through an educated machine that includes machine learning and environmental modeling that considers boundaries, pollution sources, obstacles, wind dynamics, and topography. Specifically, we present a combined ML-based linear regression model designed to infer dense concentration maps from a sensor array using state-of-the-art simulation methods. We dub this the Ridiculously Simple data-driven air pollution Interpolation Method (RSIM). The RSIM method was evaluated on both synthetic and real-life-based simulation models. The synthetic scenarios included an industrial area with point pollution sources and an urban road surrounded by buildings simulating traffic-related pollution. The real-world environment consisted of sensor data and simulations from Antwerp, Belgium. The results indicate that this method outperforms standard techniques for reconstructing dense pollution maps from sparse sensing, and demonstrates significant promise for other real-world applications.

[*] Corresponding Author: B. Fishbain - fishbain@technion.ac.il

Keywords: Air pollution interpolation, Machine Learning, Simulated-data

1. Introduction

Air pollution presents significant health risks that include respiratory diseases, cardiovascular ailments, and premature mortality. As an environmental hazard, it is related to acid rain and global warming, which adversely affect ecosystems and biodiversity (Molina and Molina, 2004; Huttunen et al., 2012; Zhang et al., 2019). The first line of defense to air pollution is air quality monitoring. However, monitoring is inherently spatially constrained (Moltchanov et al., 2015; Castell et al., 2017). Interpolation methods are generally employed to address these spatial limitations. These methods utilize sparse sensor data to generate dense pollution maps. The complexity of interpolation increases in areas where pollution levels fluctuate due to factors such as traffic density, industrial emissions, airflow boundaries, and topography.

Broadly speaking, interpolation methods for air pollution can be classified into mathematical, physical, geo-statistical, and Machine Learning (ML) models. Mathematical methods assume some smoothness or variation constraints. Physical methods build on Chemical Transport Models (CTMs) that apply predefined rules to estimate pollutant levels in locations with no sensor (Terrenoire et al., 2015; Petetin et al., 2016; Menut et al., 2013; Deligiorgi et al., 2011; Nebenzal and Fishbain, 2018; Nebenzal et al., 2020; Oettl, 2015). Geo-statistical methods account for the spatial correlation of pollutant levels across different locations using statistical techniques to model spatial variations and estimate levels in areas without sensors (Chang, 2022; Berman et al., 2019; Wu et al., 2018). Machine Learning (ML) methods leverage algorithms that are trained on existing data and use them to estimate pollutant levels in areas without measurements (Zhou et al., 2018; Hu et al., 2017; Ordieres et al., 2005).

Although they have been studied extensively, all these air pollution interpolation methods suffer from inherent limitations. CTM methods assume some constraints on the observations, such as the smoothness of the pollution field (Holmes and Morawska, 2006). To account for some of these limitations, interpolation methods based on locating the pollution sources have been suggested (Nebenzal and Fishbain, 2017; Nebenzal et al., 2020). In these methods, an inverse CTM transform, inspired by the Hough transform (Ballard, 1981), finds the source locations. Then, once these sources have been located, the same CTM is used to infer the entire dense pollution map. All CTM-based models are accurate in simple environments that contain no obstacles such as buildings, and in situations where the source configuration is straightforward, such as a set of one or more point sources (i.e., factory) or a few line sources (i.e., roads). By contrast, CTMs fall short in complex environments with non-trivial emission patterns. Geo-statistical methods are excellent for large-scale coarse estimation, but they fail to account for the variability of air pollution levels in small neighborhoods (Terrenoire et al., 2015).The main drawback of ML methods is their need for large amounts of data that must represent all possible states of the observed system (Nogueira et al., 2017). This makes data availability a crucial element that often precludes the use of ML methods, thus making it hard to assess the accuracy of these methods in real-world environments.

This paper presents a *Ridiculously Simple data driven air pollution Interpolation Method (RSIM) educated machine* approach (Kendler et al., 2022; Geltman et al., 2024) to overcome these limitations that works by integrating ML algorithms with a Digital Twin (DT) that emulates the observed environment. The integration of ML and DT alleviates the need for large datasets, reduces the model's training time as compared to classical ML methodologies, and allows for simpler neural network architectures. In a DT, a virtual representation of a system is modeled for both the experiments and the analysis (Thelen et al., 2022). DTs have been used in various fields and industrial sectors including civil and environmental engineering (Todorov and Dimov, 2023). Running simulations on an environment's digital twin can compensate for the lack of available data because the interpolation is based on air pollution patterns generated by the DT. Once the DT generates simulated data, a model is trained to reconstruct the complete, dense pollution map from a set of sparse samples extracted from the twin's output. These sparse samples correspond to sensor measurements in the real world. We demonstrate the applicability of this method in two synthetic and one real-life environment. The synthetic environments were generated by GRAMM GRAL software for air pollution and wind simulations (GRAL, 2024). The real-world application used real sensor data and dense simulated air pollution maps of varying conditions in Antwerp, Belgium that took hourly reported wind patterns, pollutant emissions and weather conditions into consideration. The reconstructed pollution maps were better quality than several other interpolation methods. The generated dense maps were also subjected to a Singular Value Decomposition (SVD) of the weights matrix, which breaks down the pollution maps into an orthogonal basis of pollution patterns. We examine the relationship between the number of sensors, the magnitude of the sensing noise, and the overall performance of the method. The results of the real-world setting show promising potential for applications.

2. Methods

2.1. Notation

The interpolation was cast as a signal reconstruction problem. The goal was to reconstruct the entire signal over Ω, from a set of sparse measurements, where all the measurements were considered to be the uncorrupted signal and all the missing data were the samples to be reconstructed. This was done by learning the dependencies between variables as they appeared in the dispersion models.

Let Ω be the region of interest and $\mathcal{F}$ be a set of scalar fields over Ω:

$$\forall f \in \mathcal{F}; \quad f : \Omega \to \mathbb{R}, \tag{1}$$

and let $\Omega^s \subset \Omega$ be a set of points in Ω, where sensors are located. The total number of sensors is denoted as $s = |\Omega^s|$. The encoding function $e(f) : \mathcal{F} \to \mathbb{R}^s$ takes a pollution map and returns a vector $\vec{z}$ of sensor readings.

$$\forall \omega \in \Omega^s; \quad z_\omega = e(f(\omega)) \tag{2}$$

The encoder $e(f)$ describes a process of sensing air pollution via a sensor array. To enable proper representation of the actual sensor's mode of operation, $e(f(\omega_i))$ takes the value f_{ω_i}, adds a Gaussian noise and the final number is rounded off. These represent the sensing noise and accuracy:

$$\forall \omega \in \Omega^s \quad z_\omega = e(f(\omega)) = round[f(\omega) \cdot (1 + \eta_\omega)]; \quad \eta_\omega \sim \mathcal{N}(0, \epsilon^2) \tag{3}$$

where η_ω is the sensor's additive white Gaussian noise at $\omega \in \Omega$ with a standard deviation of ϵ. The number of sensors (s) and their precision as described by a noise factor (ϵ) could vary. The sensitivity analysis for both of these parameters is described in the results section.

Using this notation, the goal is to find a reconstruction function (decoder), d, that estimates the pollution map $\hat{f}$:

$$d : \mathbb{R}^s \to \mathcal{F}; \quad \hat{f} = d(\vec{z}) = d\Big(e(f)\Big) \tag{4}$$

Figure 1 visualizes the concept of equation (4). The encoder in this scheme is the sensor measurements, whereas the encoder is the process of creating the desired output; i.e., an interpolation and extrapolation method that reconstructs the original pollution map based on sparse measurements in space (equation 4).

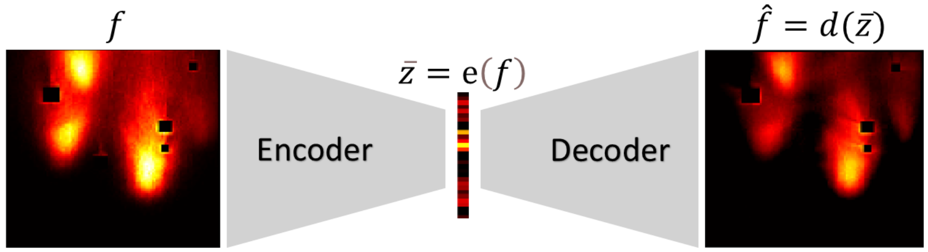


Figure 1: Encoder - Decoder flow

Since there are infinite ways to choose d we observed a class of such functions D. Each $d \in D$ took a sensor vector $\vec{z}$ and returned an estimation of the air pollution map $\hat{f}$ on ω. For every $d \in D$ we could thus estimate whether the reconstruction was satisfactory when $\hat{f}$ was compared to the ground-truth pollution map f by some distance measure, i.e., a loss function $l(f, \hat{f})$. The optimal decoding function, $d^* \in D$, was the one that provided the minimum loss over Ω.

Let us denote $P_{\mathcal{F}}$ as the probability density of $\mathcal{F}$ as it appears on Ω. The probability density captures the spatiotemporal dependencies,the physical bounds and the rules of fluid dynamics of the pollution. Thus, to find an optimal reconstruction function $d^* \in D$, we need to minimize the expectancy of the loss function over the entire probability density function$P_{\mathcal{F}}$:

$$d^* = \arg\min_{d \in D} \mathbb{E}_{f \sim P_{\mathcal{F}}}[l(f, \hat{f})] \tag{5}$$

To solve this optimization problem, a collection of m observed functions $f_1, ..., f_m \in \mathcal{F}$ was used to analyze the probability density $P_{\mathcal{F}}$ through either observation or simulation. The underlying assumption was that the simulations captured the true distribution of the air pollutants in the region of interest. This shifted the interpolation problem to the generative ML domain, where the database consisted of the m observed functions. The loss function was set to the L^2 norm, the Mean Squared Error (MSE). Therefore, the optimization problem became:

$$d^* = \arg\min_{d\in D} \frac{1}{m}\sum_{j=1}^{m} \|f_j - \hat{f}_j\|^2 = \arg\min_{d\in D} \frac{1}{m}\sum_{j=1}^{m} \|f_j - d\Big(e(f_j)\Big)\|^2 \tag{6}$$

2.2. Case studies

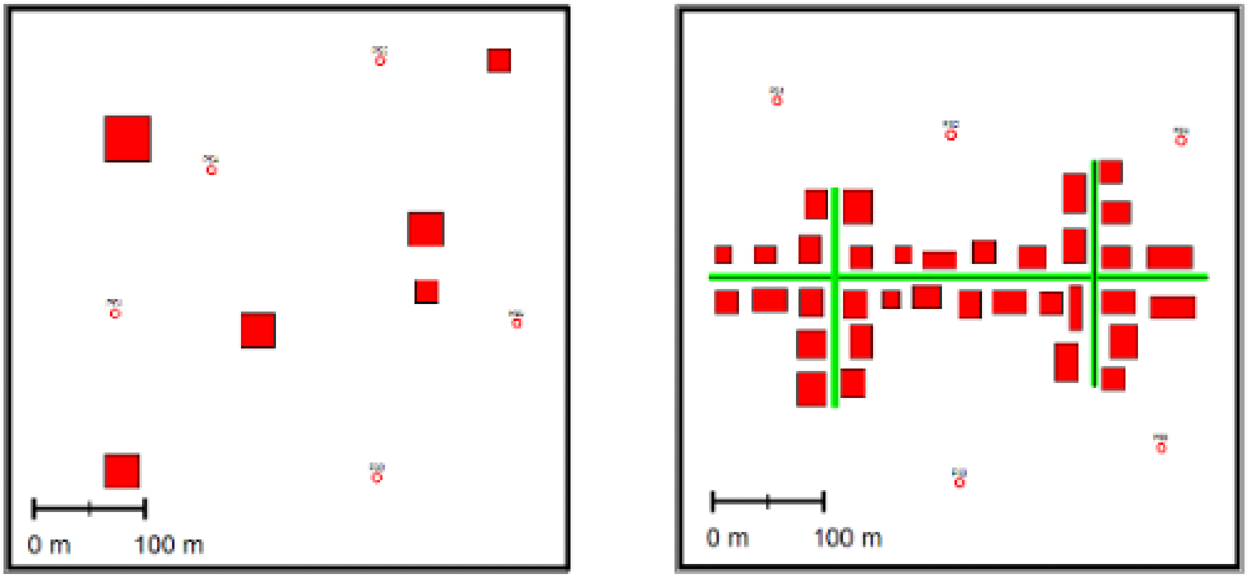
We used both synthetic and real-world data to develop and validate the method. The synthetic data consisted of two different scenarios and the real-world data provide a practical example.

2.2.1. Synthetic Scenarios

The synthetic scenarios were created to simulate different environmental conditions and pollution source configurations. These scenarios allowed for controlled experimentation and sensitivity analyses. These two simulations were presented in Mano et al. (2022), where the original goal was to find an optimal set of sensor locations using information theory. The main finding was that sensors should be located in locations with high entropy. In this work, for the synthetic scenarios, the sensor locations were based on Mano et al. (2022). The simulations were created using the Graz Lagrangian Model - GRAL (Berchet et al., 2017), an atmospheric dispersion modeling tool developed by the Institute of Meteorology at the University of Natural Resources and Life Sciences, Vienna (GRAL, 2024). The simulations were conducted for the region of interest with a fixed geometry and included various typical air pollution events. The following synthetic configurations were investigated:

- **Setting 1: Industrial Area** - This setting corresponds to a typical industrial environment with five point sources of pollution that appear as red circles on the map, and a few buildings indicated by the red squares that act as obstacles. This scenario is depicted in Figure 2a.
- **Setting 2: Urban Road** - This setting simulates an urban environment with linear pollution sources; i.e., roads indicated by the green lines between rows of dense buildings. This configuration is depicted in Figure 2b.

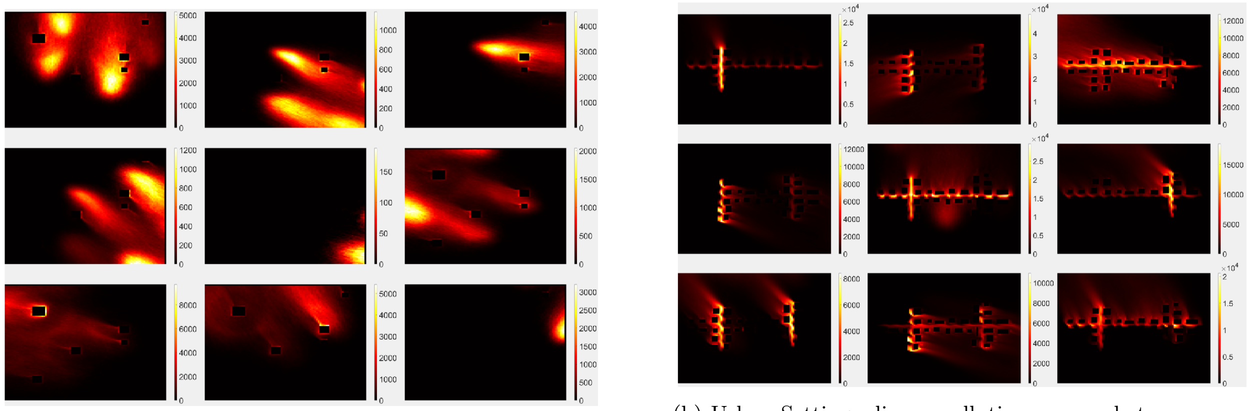
Based on the above notation, Ω was divided into a regular grid of 100×100 catchments; i.e., $\{\omega\} \in \Omega$. The assumption was that the catchments were small enough for the pollution to be constant all over the catchment, and that a single measurement at a given point in time would represent the pollution level across the catchment. The pollution maps were written as $\mathcal{F} = \mathbb{R}^{100\times100}$. The goal was to restore the full matrix f using the sensor array $\vec{z}$. During training and the comparison to other methods, the number of sensors was set to $s = 30$; however, for the sensitivity analysis s ranged from 1 to 30. Within each hypothetical setting, a total of 30,000 pollution maps were generated, where for each realization, the emission rates and wind speed and direction varied. More details are provided in Mano et al. (2022). Figure 3 shows 9 different realizations of each GRAL configuration. The dense pollution maps are color-coded with black as zero, through red to yellow. A color bar is provided for each realization map since each had it own dynamic range.



(a) Setting 1 - point sources with few buildings (Industrial Area)

(b) Setting 2 - linear pollution sources between rows of buildings (urban road)

Figure 2: GRAL synthetic air pollution configurations



(a) Industrial Setting - point sources with a few buildings

(b) Urban Setting - linear pollution sources between rows of buildings

Figure 3: Examples of pollution maps generated by a simulator so that the ML model could mirror their patterns.

2.2.2. Real-world scenario

The application of the method to a real-world environment used the city of Antwerp, Belgium, as an instance of a complex urban environment with diverse pollution sources and considerable spatial variability in air quality. The city is characterized by a dense building infrastructure, a flat yet varied landscape, and a mix of roads, vegetation, and open spaces. The major pollution sources in the city include vehicular emissions from its busy road network, industrial activities concentrated in the port area, and residential heating. These features influence pollutant dispersion, resulting in highly heterogeneous pollution patterns (De Craemer et al., 2020). For purposes of analysis, high-resolution pollution maps generated by ATMOSYS were utilized. This high-resolution air quality modeling system was developed by the European LIFE program, and is known for its ability to simulate pollutant concentrations based on emission inventories, meteorological data, and dispersion patterns. These maps are displayed as a 1100 x 1200 grid, providing detailed spatial representations of three key pollutants: $\mathrm{PM}_{2.5}$, PM_{10}, and ozone. The pollution maps reflect the average hourly concentrations and incorporate factors such as meteorology, urban morphology, and emissions from traffic and industrial activities, thus making them a reliable dataset to validate the method. Figure 4 depicts pollution maps from this dataset that have similar structural patterns for different pollutants. This underscores the effectiveness of the method that can leverage variations in pollutant concentrations in low-dimensional space. The ozone maps, however, show the reverse pattern from $\mathrm{PM}_{2.5}$ and PM_{10} due to the inverse relationship between ozone and particulate matter. This is due to chemical processes in the atmosphere, where there is often less ozone formation in areas with high particulate concentrations, such as near traffic or industrial zones. This further supports the value of the low-dimensional nature of the data representation, and makes the interpolation approach particularly suitable for these settings.

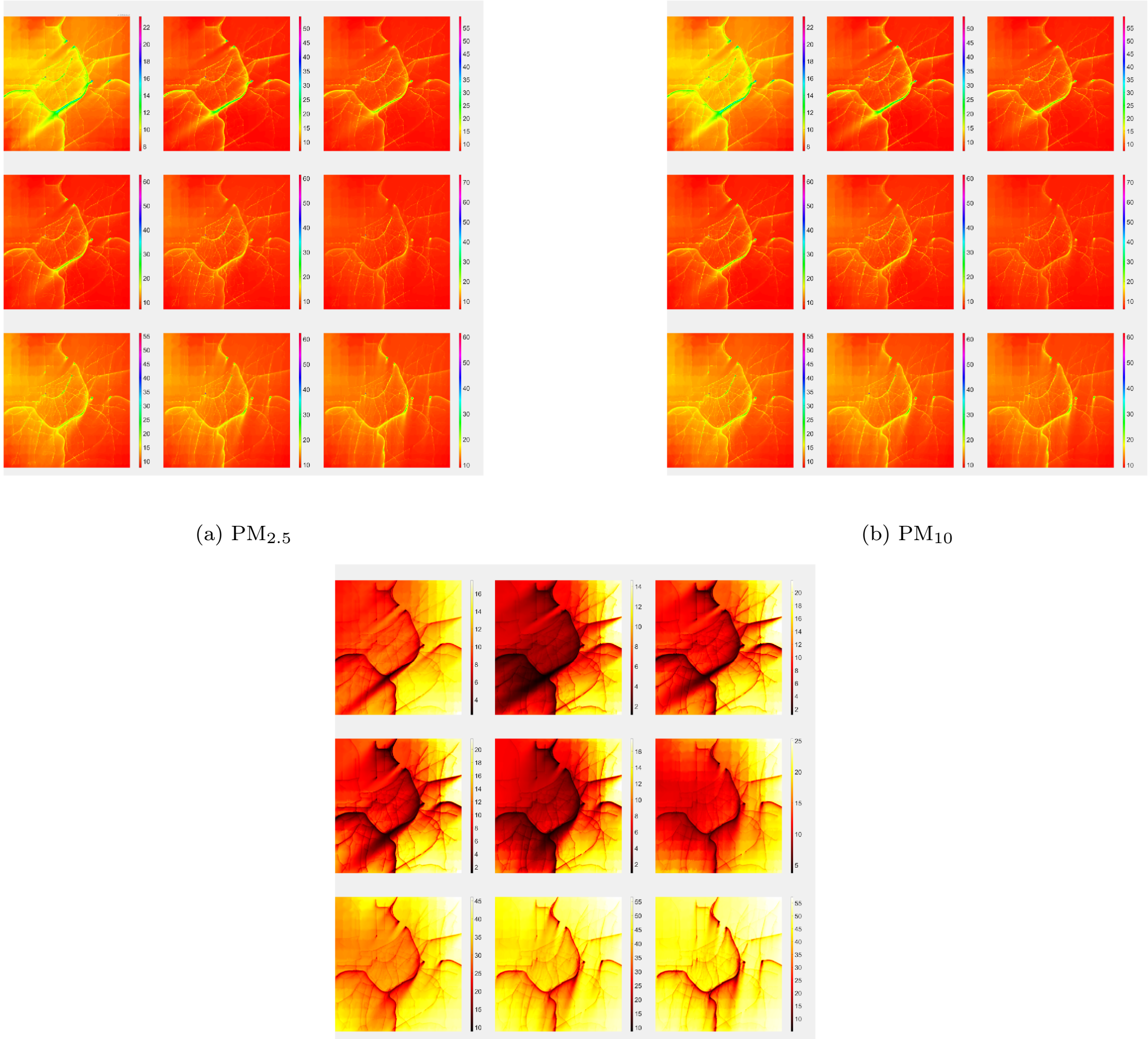
In addition to the simulated maps, the analysis incorporated real-world sensor readings of $\mathrm{PM}_{2.5}$, PM_{10}, and ozone concentrations from Antwerp's air quality monitoring network. In total there were 33 sensors s, and 1794, 1793 and 2257 samples m for $\mathrm{PM}_{2.5}$, PM_{10}, and ozone, respectively. Further details on the study design and measurement instruments can be found in Van Poppel et al. (2023). The sensor dataset is publicly available at Yatkin et al. (2022).

The domain of interest was defined as $\Omega = |1100 \times 1200|$, and the pollution maps took the form of $f \in \mathbb{R}^{1100\times 1200}$. The aim was to restore the entire matrix f for each point in time, using the sensor reading vector $\vec{z}$ at this time. Hence, the database consisted of pairs $\forall j \in 1, \cdots, m; [\vec{z}_j, f_j]$. Note that there was no encoding in this setting because the sensor readings were not generated from the simulated data, but rather from a real sensor array. Thus, the optimization function was as follows:

$$d^* = \arg\min_{d\in D} \frac{1}{m}\sum_{j=1}^{m} \|f_j - d(\vec{z}_j)\|^2 \tag{7}$$

2.3. The RSIM model

Here, for the class of functions $\mathcal{D}$, a linear regression model was implemented because it provided better results and had lower computational complexity. In this linear model, for a



(a) PM$_{2.5}$

(b) PM$_{10}$

(c) Ozone

Figure 4: Examples of pollution maps generated by a simulator so that the ML model could mirror their patterns.

given location in $\omega \in \Omega$, the pollution level $\hat{f}(\omega)$ was estimated using a weight vector W_ω in the following form:

$$\hat{f}(\omega) = \langle W_\omega, \vec{z} \rangle + b_\omega \tag{8}$$

This resembles the mathematical logic of both the Inverse Distance Weights (IDW) (Buteau et al., 2017) and Kriging (Jeong et al., 2005) methods. In IDW, the weights are determined by distance, and in Kriging by statistical resemblance. In the current study the weights were learned from the data. The training process was crafted to fit both linear and non-linear models and used a neural network training scheme. For the linear model, the network had a single layer without a non-linear activation function. This made it possible to test many configurations, such as increased noise magnitude, different numbers of sensors, shifts between the three settings, choice of hyper-parameters for training, and different model architectures (for non-linear models such as deep neural networks).

2.3.1. RSIM model analysis with SVD

Because the reconstruction function can be represented as a linear regression model, it can basically be seen as an affine transformation in a finite-dimensional space. This transformation can be characterized by a weight matrix $W \in \mathbb{R}^{|\Omega| \times s}$ and a bias vector $\bar{b} \in \mathbb{R}^{|\Omega|}$. Analyzing the weight matrix W with linear algebraic methods generated significant insights. Specifically, the i-th column of W corresponded to the transformation of the i-th standard basis vector in $\mathbb{R}^s$, which could be interpreted as a gradient map indicating the change in each component of the input vector. This gradient map reflected the impact of a unit increase in the value of a specific sensor on the resulting pollution map. Singular Value Decomposition (SVD) on the matrix W, expressed as $W = U\Sigma V^T$, resulted in a unitary real matrix U and a rectangular diagonal matrix Σ. The columns of U associated with the non-zero singular values in Σ formed an orthogonal basis that defined the span of matrix W that provided a visual representation of the distribution and directional characteristics of the pollution. The singular values quantified the relative importance of these vectors in reconstructing the pollution map. Thus overall, the columns of W and the corresponding non-zero columns of U yielded distinct visual interpretations of the reconstruction process.

2.3.2. RSIM training

The model was trained in a MATLAB coding environment. Training in the framework of a neural network employed the Adam optimizer (Kingma and Ba, 2015) as the gradient descent optimization method. This process spanned 10 epochs, with a batch size set to 128. The initial learning rate was established at 0.1, which was reduced tenfold after each epoch. Regularization was applied with a coefficient value of 5. The datasets for training, validation, and testing implemented a random split, allocating 70%, 10%, and 20% for training, testing and validation respectively. For the sensitivity analysis (conducted solely with the synthetic data), a multiplicative noise level of 0.1, 0.3, 0.5, 0.7, and 0.9 levels with respect to the mean signal's level were used. The analysis encompassed different numbers of sensors, from 1 to 30.

2.4. Evaluation methods

To gauge the efficacy of the model and to compare it to various interpolation methods, a set of benchmarks, which included both qualitative and quantitative assessments, was employed:

- **Visualization**: the outputs of the model and the interpolation methods were visually represented as pollution maps.
- **Mean Root Mean Square Error** (MRMSE): the average magnitude of the difference between the predicted value and the ground truth. Low RMSE indicates better agreement between predictions and actual values.
- **Mean Correlation Coefficients** (MCorr): the level of the two-dimensional linear correlation between the predicted values and the ground truth. High MCorr signifies a stronger alignment between predictions and ground truth.

The comparative analysis encompassed a range of interpolation methods, each evaluated against the model's performance. These interpolation techniques were chosen to assess the model's predictive capabilities against established benchmarks. The following interpolation methods were employed for comparison:

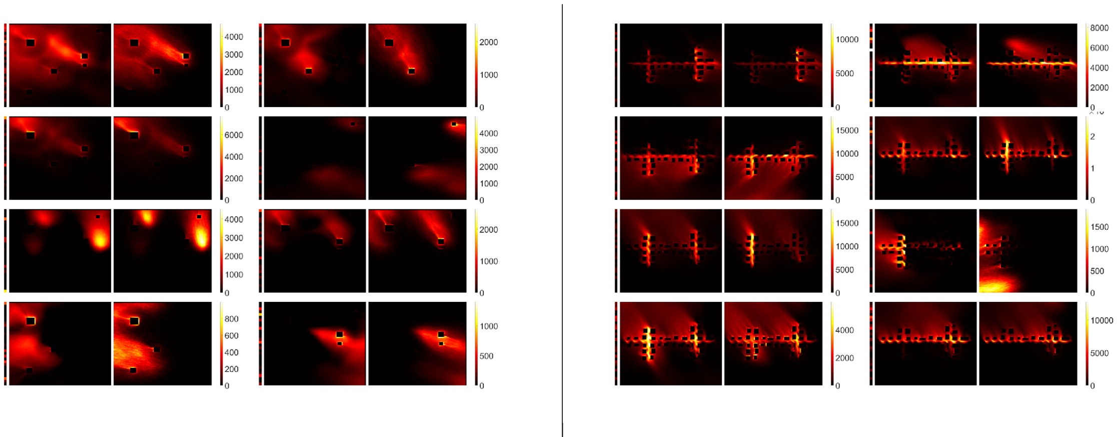
- **Linear Interpolation**: estimates values by assuming a linear relationship between known data points, resulting in a straight line interpolation.
- **Cubic Interpolation:** uses a cubic polynomial to interpolate between data points, providing continuous second derivatives, guaranteeing a physically smooth estimation with reduced oscillations.
- **Nearest Neighbor Interpolation**: assigns the value of the nearest data point to an unobserved location, resulting in a piecewise constant estimation.
- **Inverse Distance Weighting (IDW):** assigns weights to nearby data points based on their inverse distances to the target location.
- **Ordinary Kriging:** a geostatistical method that considers both the spatial correlation and the underlying trend of the data for interpolation.
- **Universal Kriging:** similar to ordinary Kriging, it also takes into account the trend and spatial correlation but also considers external drift terms.

3. Results

3.1. Synthetic scenarios

3.1.1. Visualization

Figure 5 presents the reconstructed pollution maps from the sensor vectors ($z(\omega)$) as described by Equation (4). The maps are color-coded from black through red to yellow. Each reconstructed dense pollution map is presented next to its corresponding original simulated maps, which serve as the ground truth. Each row displays two separate reconstruction examples side by side. Each example presents (from left to right): the sensor vector from which the pollution map was reconstructed, the reconstructed pollution map, the ground truth map, and a color bar with the pollution units. Note that each row presents a different colormap scale.



(a) Setting 1 - Industrial area

(b) Setting 2 - Urban road

Figure 5: Examples of pollution maps reconstructed by the model from sparse sensor readings. Each row lists (from left to right): the sensor measurement vector, the reconstructed pollution map, the ground truth pollution map, the color bar for pollution units

Figure 6 depicts one example of the dense pollution maps generated for each synthetic environment with RSIM and the interpolation methods described in section 2.4. Figure 6 clearly shows that the RSIM method successfully inferred the spatial dependencies, as compared to the-state-of-the-art used for comparison. In particular, RSIM was able to preserve the intricate patterns present in the original map.

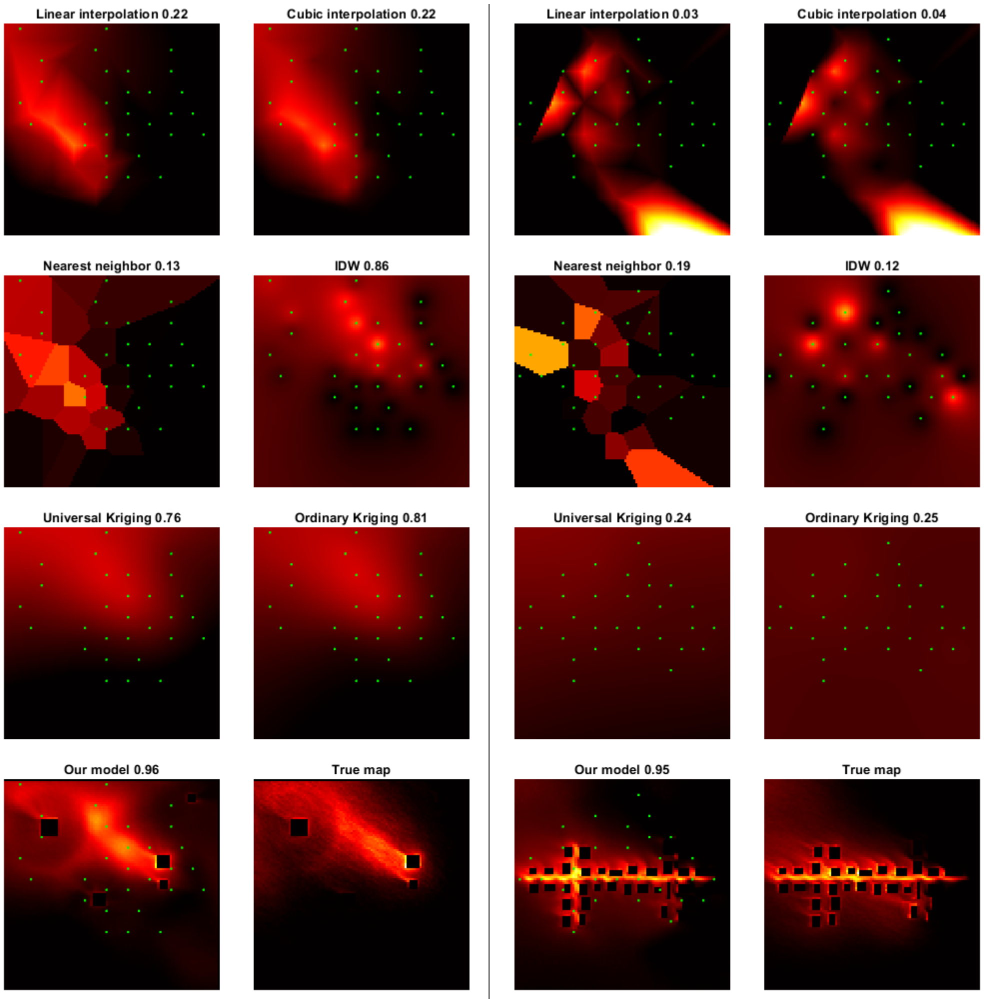
3.1.2. Benchmarks

Table 1 presents the benchmark scores for the entire set of synthetic datasets for each interpolation method. Note that each interpolation method read the same input; i.e., a sensor vector, z_ω with added noise (Equation (3)) and created a pollution map.

3.1.3. Sensitivity Analysis

After the mathematical formulation of equations (3)–(6) the resulting encoding process had two parameters: the number of sensors s, and the noise factor η_ω. Both parameters capture the resources available for monitoring; namely, the amount and the quality of the sensor array. Figure 7 plots the MCorr and RMSE for different noise levels against the number of sensors, s. Figure 7(a) and (c) present the MCorr for the industrial and urban synthetic scenarios respectively. Figures (b) and (d) present the RMSE for these two scenarios. The noise level is governed by the noise multiplier magnitude ϵ (see Equation (3).

As expected, as the number of sensors increased, their quality improved (i.e.,lower noise factor), and the reconstruction performance improved. A knee point was observed at approximately 10 sensors in the industrial setting, and 5 sensors in the urban setting. Beyond the knee point, the marginal benefit of adding more sensors was negligible. At that knee point, high-quality sensors with up to a 0.1 noise factor yielded results comparable to 30 low-quality sensors plagued by 0.9 noise. This insight has implications for monitoring policies and resource allocation.



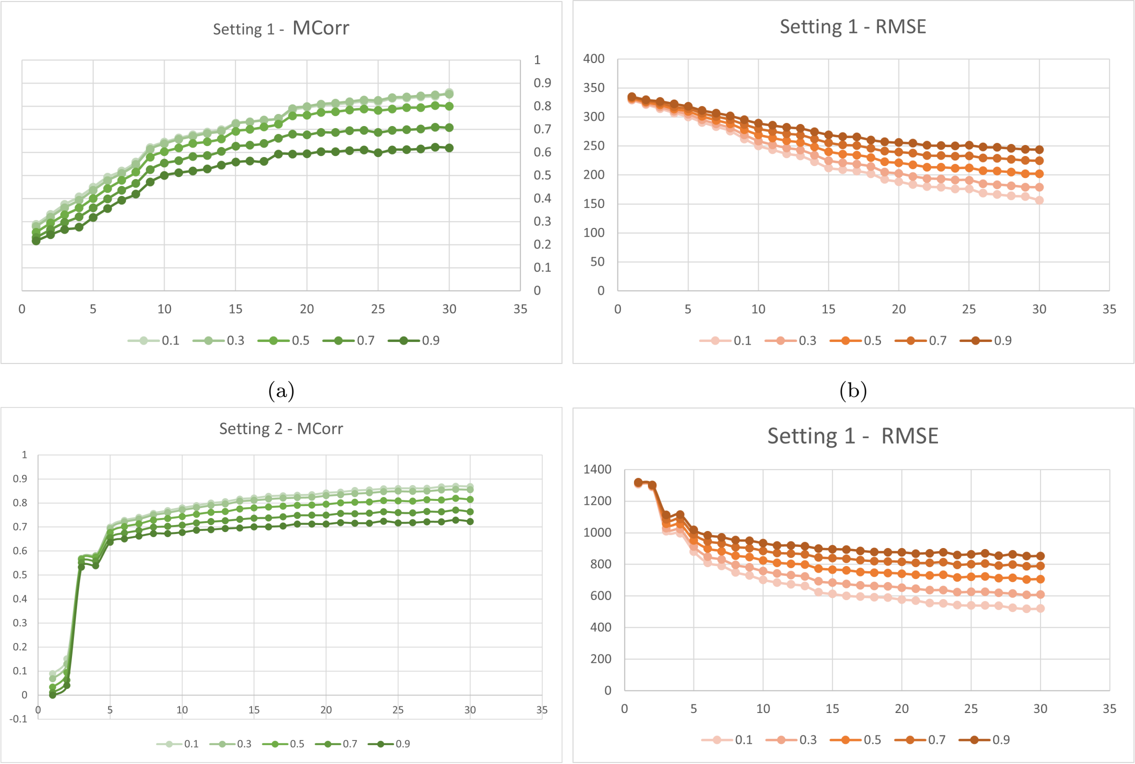
(a) Setting 1

(b) Setting 2

Figure 6: Samples of pollution maps.

Table 1: MRMSE and MCorr benchmark results for the synthetic scenarios in all interpolation methods

	Method	MRMSE	MCorr
Industrial area	Linear	471.9	0.17
	Cubic	470.4	0.17
	Nearest neighbor	417.6	0.19
	IDW	233.4	0.71
	Universal Kriging	293.4	0.58
	Ordinary Kriging	278.0	0.51
	RSIM	**156.5**	**0.86**
Urban road	Linear	2604.1	0.08
	Cubic	2595.4	0.09
	Nearest neighbor	1857.7	0.03
	IDW	1409.0	0.02
	Universal Kriging	1326.0	0.07
	Ordinary Kriging	1284.4	0.01
	RSIM	**520.2**	**0.87**



(a)

(b)

(c)

(d)

Figure 7: Sensitivity Analysis of the mean correlation (MCorr), Figures (a) and (c), and the RMSE (b) and (d), for the synthetic scenarios, industrial (a) and (b), and urban, (c) and (d) settings

3.1.4. Weight matrix decomposition

As described in the Methods section, the reconstruction function was represented by a weight matrix $W \in \mathbb{R}^{|\Omega| \times s}$ and a bias vector $\bar{b} \in \mathbb{R}^{|\Omega|}$. The i-th column of the weight matrix W represented the image of the gradient map with respect to the i-th sensor. The typical assumption was that for each

$bar\omega \in \Omega$, the effect on the pollution level at $\bar{\omega}$, $f(\bar{\omega})$, would decrease as the distance between $\bar{\omega}$ and this specific sensor increased. Hence, the weight $W_{\bar{\omega},i}$ would decrease with the distance of $\bar{\omega}$ from the sensor. However, this was not always the case, since these sensors were not statistically independent. Figure 8 shows these gradient maps with the same color coding as above. The green dots indicate the sensor locations.

In the industrial setting, the sensor gradient maps presented Gaussian-like curves around the sensors with edge effects around obstacles. This was expected since it resembles other air pollution interpolation methods such as IDW. However, in the urban setting these gradient maps emerged as less intuitive. Specifically, some sensors only monitored the inner part of the street while others were impacted by the background, some were close to the affected area while others were farther away. This was probably due to the complicated wind regime in this setting.

Applying SVD to the weight matrix $W = U\Sigma V^T$, on the first s columns of U provided the orthogonal basis of the pollution patterns. Their significance in descending order matched the singular values $\sigma_i = (\Sigma)_{ii}$. Figure 9 depicts these values for each synthetic scenario. The maps that correspond to higher singular values provide a striking depiction of the pollution sources. In both settings, as the singular value decreased, the patterns exhibited greater variance. This can be ascribed to variations in finer details across the pollution maps. However, there were overarching pollution patterns that remained representative across all the data. Each restored pollution map consisted of a linear combination of these patterns, with both positive and negative scalars; as shown by the inverse color scheme.

3.2. Real-world Environment

To evaluate the performance in a real-world setting, a series of visualizations is presented in Figure 10 comparing the reconstructed pollution maps generated by the model to the ground truth maps. These visualizations include examples for the three types of pollutants $\text{PM}_{2.5}$, PM_{10}, and ozone, thus providing a clear comparison of the model's outputs as compared to the reference maps. The reconstructed maps demonstrate the model's ability to replicate the spatial distribution of pollutants with high fidelity by capturing areas of elevated concentrations and regions with lower pollution levels.

3.2.1. Benchmarks

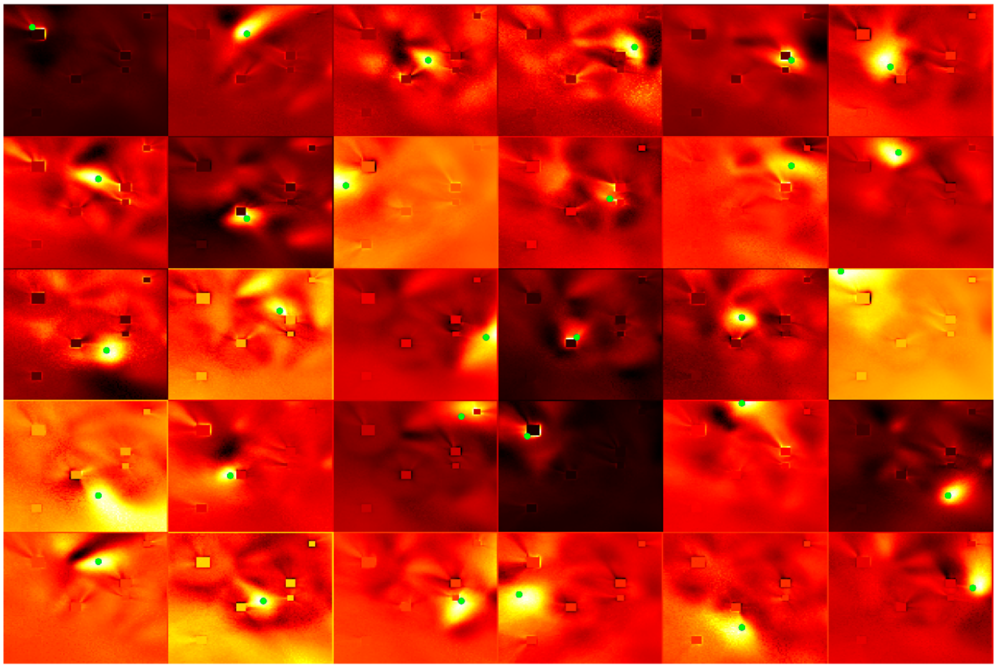
For the real-world data, the benchmarks were calculated and averaged in a similar fashion as for the synthetic scenarios. The mean performance scores for the models are summarized in Table 2.

4. Discussion

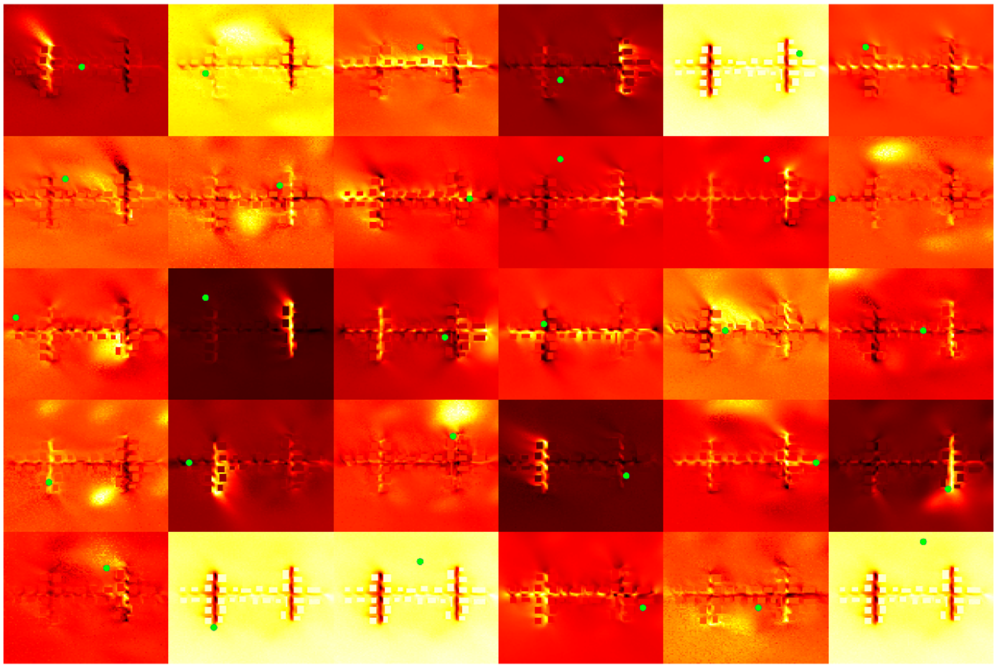
The outcomes of this investigation underscore the benefits and potential applications of incorporating simulated data into machine learning techniques for air pollution interpolation.

Table 2: Results

Pollutant	**MRMSE**	**MCorr**
$PM_{2.5}$	6.33	0.77
PM_{10}	6.16	0.76
Ozone	12.43	0.77

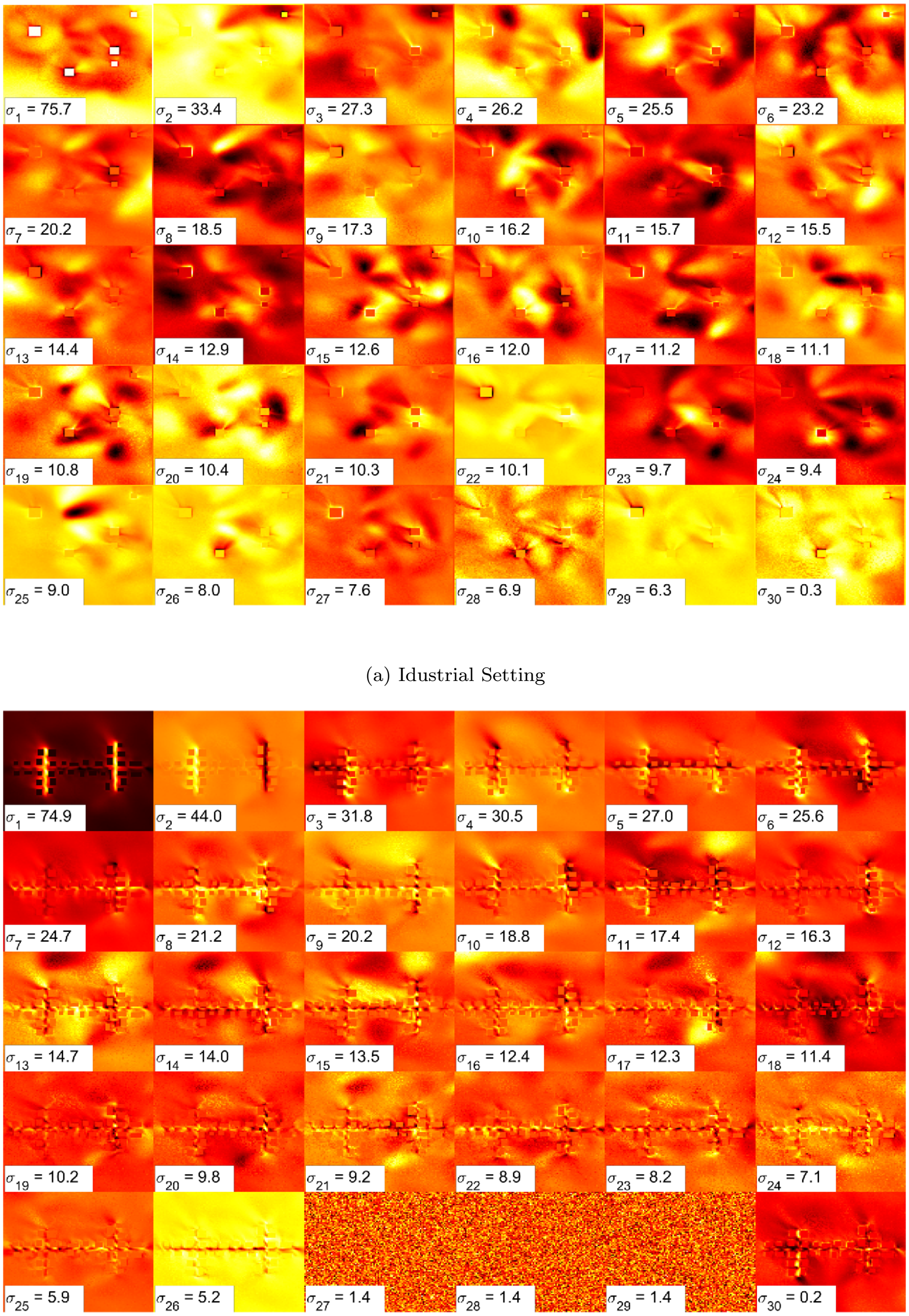


(a) Industrial Area



(b) Urban area

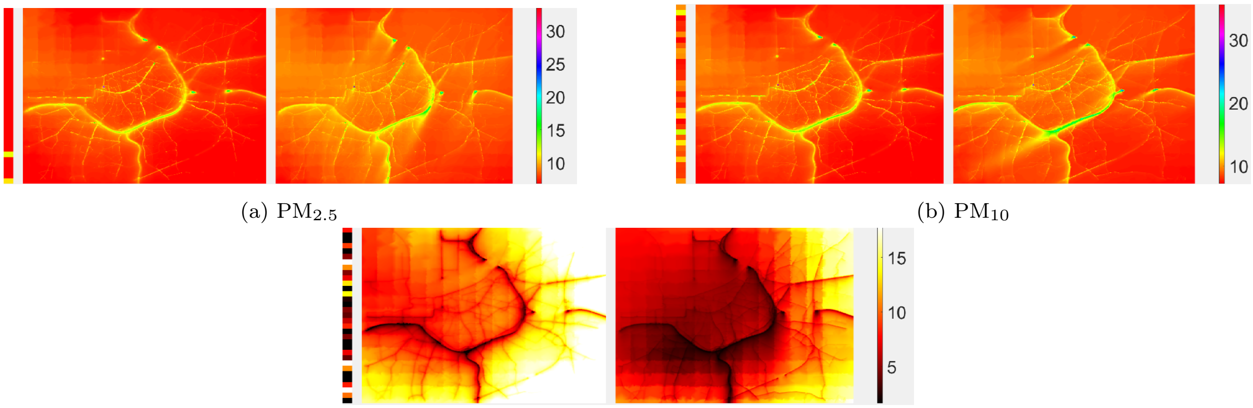
Figure 8: Gradient maps for each sensor



(a) Idustrial Setting

(b) Urban Setting

Figure 9: SVD basis of pollution patterns



(a) PM$_{2.5}$

(b) PM$_{10}$

(c) Ozone

Figure 10: Samples of the test set results. Left to right: sensor vector input, model prediction, true map, color bar.

State-of-the-art methodologies are constrained in their ability to accurately capture the complex spatial variability of pollutant dispersion, particularly in urban settings with diverse sources of pollution and obstacles such as buildings. By employing high-resolution digital twin simulations of the environment, the methodology proposed here was shown to overcome these limitations by effectively reconstructing pollution maps with heightened accuracy and detail.

The empirical experiment utilizing data from Antwerp illustrates the practicality of this approach. The concordance between the reconstructed maps and the ground truth simulations indicates that the model can capture fine-grained patterns of pollutant dispersion, even when hampered by noisy and sparse sensor data. This capability is paramount for urban air quality management, where access to dense and accurate pollution maps directly impacts public health policies and mitigation strategies.

In this study, an innovative encoder-decoder methodology was utilized to address the challenge of reconstructing detailed pollution maps from sparse sensor data. In this case, the encoding process corresponded to the sensing process itself, where the available sensor measurements, typically sparse and noisy, served as the input. These sensor readings, although limited, captured essential environmental information. The decoder then reconstructed the comprehensive pollution map, using this encoded information to generate a detailed distribution of pollutants across the entire area.

The work presented here underscores the adaptability and efficiency of the Kendler et al. (2022) Educated Machine framework. Unlike traditional approaches that rely on labor-intensive labeling of large datasets for each scenario, the EM leverages a physical model and measurable information to deliver accurate solutions across various domains. These two foundational components eliminate the need for extensive memorization, thus aligning with the features of human cognition. For instance, a skilled mechanic can repair a range of tools by applying general mechanical knowledge—analogous to the physical model—and by analyzing the specific characteristics of each instrument; i.e., the measurable information. Kendler and Fishbain demonstrated the EM's capabilities in the case of the non-linear mixing of reflectance spectra between target and background materials. In that context, the properties of the mixing model and the reflectance spectrum of the target material were embedded within the EM as part of its physical model and parameters. The reflectance spectra of the background materials, which exhibited substantial variability and were unpredictable up to that point, were treated as measurable information, facilitating the creation of a training dataset.

In this work, a similar approach was applied. The Lagrangian gas transport model served as the physical model, with its parameters defined by terrain properties such as buildings, roads, and topography. The measurable information comprised the chemical sensor readings for a specific case. This combination permitted the resulting EM to accurately interpolate gas concentrations across the terrains, even in areas lacking labeled measurements.

However, this approach also has shortcomings. The reliance on accurate and detailed simulations means that the quality of the results is contingent on the fidelity of the digital twin and the simulation model. Any discrepancies between the simulated environment and the actual urban landscape, such as missing sources or inaccuracies in boundary conditions, can affect the performance of the interpolation model. Furthermore, the methodology assumes static conditions during the interpolation process, which may not fully capture the dynamic nature of air pollution in rapidly changing environments.

5. Conclusion

This study introduced a data-driven approach for interpolating air pollution levels that leveraged simulated data generated from a digital environmental twin. The findings demonstrated its efficacy in both synthetic and real-world scenarios by addressing the limitations inherent to traditional interpolation techniques. By integrating the complex spatial and temporal patterns of pollutant dispersion captured in simulations, the proposed approach successfully generated accurate reconstructions of high-resolution pollution maps from sparse sensor data. The real-world evaluation further validated the method's applicability in practical urban environments. By conceptualizing the sensing process as encoding, the model demonstrated its ability to learn from limited data and fill in the gaps, thereby providing a robust solution for air pollution interpolation in urban settings.

Future research can build on these findings by exploring additional pollutant types and applying this method to other cities or regions to further generalize its applicability. Whereas this study utilized settings within a two-dimensional space, future work could expand to vertical spatial and temporal dimensions. Successful interpolation and extrapolation in such spaces would have significant value: vertical pollution patterns are typically unaddressed and unknown, and extrapolation along the temporal axis serves as a forecast. Although the current study focused on air pollution, the concepts presented here are likely to be applicable to other domains where simulated data are more readily available than real-world measurements.

Acknowledgements

This work was partially funded by the Israel Water Authority and the Israel Ministry of Science and Technology. The authors would like to acknowledge the contribution of Michel Gerboles and Phillipe Thunis of the European Commission - Joint Research Center, and Stijn Janssen of VITO - Flemish Institute for Technological Research in making the sensor data available.

Software and Data Availability

- Name of software: **Ridiculously Simple Data Driven Air Pollution Interpolation Method**
- Developer: Alon Feldman
- Contact: alonfeldman@campus.technion.ac.il
- Date first available: Feb. 15^{th}, 2025.
- Software Required: Matlab installation
- Program language: Matlab script
- Source code at: `https://github.com/...`

- Documentation: Detailed documentation for application, data, testing, and deployment can be found at `https://github.com/...`
- Data required for local installation and use of software can be found under `https://github.com/....`

Author credit statement

Alon Feldman: Writing - Original Draft, Software, Investigation, Visualization, Formal analysis. **Shai Kendler**: Supervision, Writing-Reviewing & Editing. **Enrico Pisoni**: Software, Editing. **Barak Fishbain**: Conceptualization, Methodology, Supervision, Writing-Reviewing & Editing. During the preparation of this work, the authors used ChatGPT in order to enhance the clarity and readability of the manuscript, assist with language refinement, and generate draft formulations of technical explanations. All intellectual content, scientific reasoning, and final editing were performed and approved by the authors.

References

D. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern recognition*, 13(2):111–122, 1981. URL `http://www.sciencedirect.com/science/article/pii/0031320381900091`.

A. Berchet, K. Zink, D. Oettl, J. Brunner, L. Emmenegger, and D. Brunner. Evaluation of high-resolution GRAMM-GRAL (v15.12/v14.8) NOx simulations over the city of Zürich, Switzerland. *Geoscientific Model Development*, 10(9):3441–3459, 2017. ISSN 19919603. doi: 10.5194/gmd-10-3441-2017. URL `https://doi.org/10.5194/gmd-10-3441-2017`.

J. D. Berman, L. Jin, M. L. Bell, and F. C. Curriero. Developing a geostatistical simulation method to inform the quantity and placement of new monitors for a follow-up air sampling campaign. *Journal of Exposure Science and Environmental Epidemiology*, 29(2):248–257, 2019. ISSN 1559064X. doi: 10.1038/s41370-018-0073-6. Publisher: Springer US.

S. Buteau, M. Hatzopoulou, D. L. Crouse, A. Smargiassi, R. T. Burnett, T. Logan, L. D. Cavellin, and M. S. Goldberg. Comparison of spatiotemporal prediction models of daily exposure of individuals to ambient nitrogen dioxide and ozone in Montreal, Canada. *Environmental Research*, 156(March):201–230, 2017. ISSN 10960953. doi: 10.1016/j.envres.2017.03.017. Publisher: Elsevier Inc. ISBN: 0013-9351.

N. Castell, F. R. F. Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday, and A. Bartonova. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? *Environment International*, 99:293–302, Feb. 2017. ISSN 18736750. doi: 10.1016/j.envint.2016.12.007. URL `http://www.sciencedirect.com/science/article/pii/S0160412016309989`. Publisher: Pergamon.

H. H. Chang. Geostatistical Methods for Modeling Environmental Exposures with Applications to Ambient Air Pollution. *Geospatial Technology for Human Well-Being and Health*, pages 37–47, Jan. 2022. Publisher: Springer International Publishing.

S. De Craemer, J. Vercauteren, F. Fierens, W. Lefebvre, and F. J. R. Meysman. Using Large-Scale NO2 Data from Citizen Science for Air-Quality Compliance and Policy Support. *Environmental Science & Technology*, 54(18):11070–11078, Sept. 2020. ISSN 0013-936X. doi: 10.1021/acs.est.0c02436. URL `https://doi.org/10.1021/acs.est.0c02436`. Publisher: American Chemical Society.

D. Deligiorgi, K. Philippopoulos, D. Deligiorgi, and K. Philippopoulos. Spatial Interpolation Methodologies in Urban Air Pollution Modeling: Application for the Greater Area of Metropolitan Athens, Greece. In *Advanced Air Pollution*. IntechOpen, Aug. 2011. ISBN 978-953-307-511-2.

A. Geltman, I. Levy, and B. Fishbain. Machine Education Approach for Generating Accurate NO_2 and PM_(2.5) Dense Pollution Maps in Israel. *Environmental Science & Technology - Air*, 2024.

R. GRAL. Home, 2024. URL `https://gral.tugraz.at/`.

N. S. Holmes and L. Morawska. A review of dispersion modelling and its application to the dispersion of particles: An overview of different dispersion models available. *Atmospheric Environment*, 40(30):5902–5928, Sept. 2006. ISSN 1352-2310. Publisher: Pergamon.

K. Hu, A. Rahman, H. Bhrugubanda, and V. Sivaraman. HazeEst: Machine Learning Based Metropolitan Air Pollution Estimation From Fixed and Mobile Sensors. *IEEE SENSORS JOURNAL*, 17(11), 2017.

K. Huttunen, T. Siponen, I. Salonen, T. Yli-Tuomi, M. Aurela, H. Dufva, R. Hillamo, E. Linkola, J. Pekkanen, A. Pennanen, A. Peters, R. O. Salonen, A. Schneider, P. Tiittanen, M.-R. Hirvonen, and T. Lanki. Low-level exposure to ambient particulate matter is associated with systemic inflammation in ischemic heart disease patients. *Environmental Research*, 116:44–51, July 2012. ISSN 1096-0953. doi: 10.1016/j.envres.2012.04.004.

S. Jeong, M. Murayama, and K. Yamamoto. Efficient Optimization Design Method Using Kriging Model. *Journal of Aircraft*, 42(2):413–420, 2005. ISSN 0021-8669. doi: 10.2514/1.6386. URL `https://arc.aiaa.org/doi/abs/10.2514/1.6386`. Publisher: American Institute of Aeronautics and Astronautics Inc.

S. Kendler, Z. Mano, R. Aharoni, R. Raich, and B. Fishbain. Hyperspectral imaging for chemicals identification: a human-inspired machine learning approach. *Scientific Reports 2022 12:1*, 12(1):1–10, Oct. 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-22468-7. URL `https://www.nature.com/articles/s41598-022-22468-7`. Publisher: Nature Publishing Group ISBN: 0123456789.

D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, 2015. arXiv: 1412.6980.

Z. Mano, S. Kendler, and B. Fishbain. Information Theory Solution Approach to the Air Pollution Sensor Location–Allocation Problem. *Sensors*, 22(10):3808, May 2022. ISSN

14248220. doi: 10.3390/s22103808. URL https://www.mdpi.com/1424-8220/22/10/3808/htm. Publisher: Multidisciplinary Digital Publishing Institute.

L. Menut, B. Bessagnet, D. Khvorostyanov, M. Beekmann, N. Blond, A. Colette, I. Coll, G. Curci, G. Foret, A. Hodzic, S. Mailler, F. Meleux, J.-L. Monge, I. Pison, G. Siour, S. Turquety, M. Valari, R. Vautard, and M. G. Vivanco. CHIMERE 2013: a model for regional atmospheric composition modelling. *Geoscientific Model Development*, 6(4):981–1028, July 2013. ISSN 1991-9603. doi: 10.5194/gmd-6-981-2013. URL http://www.geosci-model-dev.net/6/981/2013/. ISBN: 1991-959X.

M. J. Molina and L. T. Molina. Megacities and Atmospheric Pollution. *Journal of the Air & Waste Management Association*, 54(6):644–680, June 2004. ISSN 1096-2247. doi: 10.1080/10473289.2004.10470936. URL https://doi.org/10.1080/10473289.2004.10470936. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10473289.2004.10470936.

S. Moltchanov, I. Levy, Y. Etzion, U. Lerner, D. M. D. Broday, and B. Fishbain. On the feasibility of measuring urban air pollution by wireless distributed sensor networks. *Science of the Total Environment*, 502:537–547, Dec. 2015. ISSN 18791026. doi: 10.1016/j.scitotenv.2014.09.059. Publisher: Elsevier B.V. ISBN: 1879-1026 (Electronic)\r0048-9697 (Linking).

A. Nebenzal and B. Fishbain. Hough-transform-based interpolation scheme for generating accurate dense spatial maps of air pollutants from sparse sensing. *IFIP Advances in Information and Communication Technology*, 507:51–60, 2017. ISSN 18684238. ISBN: 9783319899343 Publisher: Springer New York LLC.

A. Nebenzal and B. Fishbain. Hough-based Interpolation Scheme for Generating Accurate Dense Spatial Maps of Air Pollutants from Sparse Sensing. *International Federation for Information Processing (IFIP) Advances in Information and Communication Technology.*, 507:51–60, 2018. ISSN 18684238. doi: 10.1007/978-3-319-89935-0_5. URL http://link.springer.com/10.1007/978-3-319-89935-0_5. ISBN: 9783319899343.

A. Nebenzal, B. Fishbain, and S. Kendler. Model-based dense air pollution maps from sparse sensing in multi-source scenarios. *Environmental Modelling and Software*, 128(April 2019): 104701, June 2020. ISSN 13648152. doi: 10.1016/j.envsoft.2020.104701. URL https://doi.org/10.1016/j.envsoft.2020.104701. Publisher: Elsevier Ltd.

K. Nogueira, O. A. B. Penatti, and J. A. dos Santos. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, 61:539–556, 2017. ISSN 00313203. doi: 10.1016/j.patcog.2016.07.001. arXiv: 1602.01517.

D. Oettl. Evaluation of the Revised Lagrangian Particle Model GRAL Against Wind-Tunnel and Field Observations in the Presence of Obstacles. *Boundary-Layer Meteorology*, 155 (2):271–287, May 2015. ISSN 15731472. doi: 10.1007/S10546-014-9993-4/FIGURES/10. URL https://link.springer.com/article/10.1007/s10546-014-9993-4. Publisher: Kluwer Academic Publishers.

J. B. Ordieres, E. P. Vergara, R. S. Capuz, and R. E. Salazar. Neural network prediction model for fine particulate matter (PM2.5) on the US–Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua). *Environmental Modelling & Software*, 20(5):547–559, May 2005. ISSN 1364-8152. doi: 10.1016/j.envsoft.2004.03.010. URL `https://www.sciencedirect.com/science/article/pii/S1364815204000830`.

H. Petetin, J. Sciare, M. Bressi, V. Gros, A. Rosso, O. Sanchez, R. Sarda-Estève, J.-E. Petit, and M. Beekmann. Assessing the ammonium nitrate formation regime in the Paris megacity and its representation in the CHIMERE model. *Atmospheric Chemistry and Physics*, 16(16):10419–10440, Aug. 2016. ISSN 1680-7316. doi: 10.5194/acp-16-10419-2016. URL `https://acp.copernicus.org/articles/16/10419/2016/`. Publisher: Copernicus GmbH.

E. Terrenoire, B. Bessagnet, L. Rouïl, F. Tognet, G. Pirovano, L. Létinois, M. Beauchamp, A. Colette, P. Thunis, M. Amann, and L. Menut. High-resolution air quality simulation over Europe with the chemistry transport model CHIMERE. *Geoscientific Model Development*, 8(1):21–42, Jan. 2015. ISSN 1991-959X. doi: 10.5194/gmd-8-21-2015. URL `https://gmd.copernicus.org/articles/8/21/2015/`. Publisher: Copernicus GmbH.

A. Thelen, X. Zhang, O. Fink, Y. Lu, S. Ghosh, B. D. Youn, M. D. Todd, S. Mahadevan, C. Hu, and Z. Hu. A comprehensive review of digital twin — part 1: modeling and twinning enabling technologies. *Structural and Multidisciplinary Optimization*, 65(12):1–55, Dec. 2022. ISSN 16151488. Publisher: Springer Science and Business Media Deutschland GmbH.

V. Todorov and I. Dimov. Unveiling the Power of Stochastic Methods: Advancements in Air Pollution Sensitivity Analysis of the Digital Twin. *Atmosphere 2023, Vol. 14, Page 1078*, 14(7):1078, June 2023. ISSN 2073-4433. Publisher: Multidisciplinary Digital Publishing Institute.

M. Van Poppel, P. Schneider, J. Peters, S. Yatkin, M. Gerboles, C. Matheeussen, A. Bartonova, S. Davila, M. Signorini, M. Vogt, F. R. Dauge, J. S. Skaar, and R. Haugen. SensEURCity: A multi-city air quality dataset collected for 2020/2021 using open low-cost sensor systems. *Scientific Data*, 10(1):322, May 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02135-w. URL `https://www.nature.com/articles/s41597-023-02135-w`.

C. D. Wu, Y. T. Zeng, and S. C. C. Lung. A hybrid kriging/land-use regression model to assess PM2.5 spatial-temporal variability. *Science of the Total Environment*, 645:1456–1464, 2018. ISSN 18791026. doi: 10.1016/j.scitotenv.2018.07.073. Publisher: Elsevier B.V.

S. Yatkin, M. Gerboles, M. V. Poppel, P. Schneider, J. Peters, C. Matheeussen, A. Bartonova, S. Davila, M. Signorini, M. Vogt, F. R. Dauge, J. S. Skaar, and R. Haugen. SensEURCity: A multi-city air quality dataset collected using networks of open low-cost sensor systems, Oct. 2022. URL `https://zenodo.org/record/7256406`.

S. Zhang, Y. Wu, H. Yan, X. Du, K. Max Zhang, X. Zheng, L. Fu, and J. Hao. Black carbon pollution for a major road in Beijing: Implications for policy interventions of the heavy-duty truck fleet. *Transportation Research Part D: Transport and Environment*, 68:110–121, Mar. 2019. ISSN 1361-9209. doi: 10.1016/j.trd.2017.07.013. URL `https://www.sciencedirect.com/science/article/pii/S1361920917300342`.

C. Zhou, K. Lin, D. Xu, L. Chen, Q. Guo, C. Sun, and X. Yang. Near infrared computer vision and neuro-fuzzy model-based feeding decision system for fish in aquaculture. *Computers and Electronics in Agriculture*, 146:114–124, Mar. 2018. ISSN 01681699. doi: 10.1016/j.compag.2018.02.006. Publisher: Elsevier.