GREENCIRC/Special Issue – 2025

AusIMM THE MINERALS INSTITUTE
IOM3 Institute of Materials, Minerals & Mining

Mineral Processing and Extractive Metallurgy
(Trans. Inst. Min. Metall. B)
1–9

DOI: 10.1177/25726641251376762
journals.sagepub.com/home/mpe
Sage

Modelling and prediction of air pollution using hybrid tree–LASSO approach

Mare Srbinovska[1], Pavel Dimovski[2], Maja Celeska Krstevska[1], Aleksandra Krkoleva Mateska[1], Vesna Andova[1], Sijche Pechkova[2] and Aleksandar Pechkov[3]

Abstract

This study employs hybrid tree–Least Absolute Shrinkage and Selection Operator approach to forecast pollutant concentrations (PM2.5, PM10, NO_2, and CO) in Skopje, using data from 2018 to 2022, which includes meteorological variables and pollution measurements from three sensor nodes. Models were trained on pre-COVID-19 data and then tested on post-COVID-19 observations to assess the pandemic's impact on air quality. Results show that models consistently overpredicted pollution levels during the pandemic, suggesting a positive effect of COVID-19 restrictions on air quality. Applications and research directions of the models in the context of metallurgy, mining, and mineral processing are discussed.

Keywords

Air pollution, pollutants, particulate matter, NO_2, CO, lasso regression, prediction

Received: 5 July 2025; accepted: 23 August 2025

Introduction

Air pollution represents a critical and escalating environmental threat with profound consequences for public health, ecosystems, and cultural heritage. Pollutants of major health concern are particulate matter (PM), ozone (O_3), nitrogen dioxide (NO_2), carbon monoxide (CO), and sulphur dioxide (SO_2). Air pollution (indoor and outdoor) causes respiratory and other diseases and is a reason for increased morbidity and mortality. According to the World Health Organization, almost all of the world's population breathes air that exceeds the WHO guideline limits and contains high levels of pollutants; low- and middle-income countries suffer from the highest exposure. Numerous studies[1,2] have linked elevated concentrations of airborne pollutants such as O_3, PM, and SO_2 with increased mortality rates and a range of respiratory and cardiovascular diseases.

The European Environment Agency (EEA) reports that in Europe, exposure to fine PM and NO_2 levels above the World Health Organization recommendations caused an estimated 239,000 and 48,000 premature deaths, respectively, in 2022.[3]

Among the most common sources of air pollution are: burning fossil fuels, industrial emissions, transportation, open burning of garbage waste, construction and demolition activities, agricultural activities, and the use of chemical and synthetic products. Meteorological conditions also play a significant role in modulating air quality. Wind speed, precipitation, and temperature influence the dispersion and concentration of pollutants. For example, higher wind velocities are generally associated with reductions in particulate levels.[4]

Efforts to mitigate air pollution span both local and global initiatives. Locally, strategies include expanding green infrastructure, transitioning to low-emission transportation, and promoting clean energy solutions. Globally, comprehensive decarbonisation efforts across the energy, transport, and industrial sectors are imperative. Green buildings and vegetated surfaces also offer additional benefits such as improved thermal insulation, reduced noise pollution, and lower energy demands.[5] Moreover, the presence of urban green spaces has been shown to mitigate PM levels,[6] pointing to the co-benefits of green infrastructure in pollution control.

The COVID-19 pandemic has unintentionally provided a natural experiment in pollution mitigation. Numerous studies have reported significant reductions in pollutant concentrations during lockdown periods.[7–9] The findings of these studies indicate that social distancing led to marked declines in PM2.5, NO_2, and CO.

In the context of North Macedonia, a country in the Western Balkans, air pollution remains a persistent concern

[1]Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje, Skopje, North Macedonia
[2]Faculty of Technology and Metallurgy, Ss. Cyril and Methodius University in Skopje, Skopje, North Macedonia
[3]Industrial Math Institute, Perth, Australia

Corresponding author:
Pavel Dimovski, Faculty of Technology and Metallurgy, Ss. Cyril and Methodius University in Skopje, Ruger Boshkovich 16, 1000 Skopje, North Macedonia.
Emails: dimovski@tmf.ukim.edu.mk; dimovski.pavel@gmail.com

driven by industrial activity, transportation, energy production, and biomass burning. The resulting health and ecological challenges necessitate targeted, evidence-based interventions. Through a rigorous analysis of historical air quality data, meteorological parameters, and other relevant environmental variables, this study aims to uncover the underlying dynamics of air pollution in North Macedonia.

To achieve this, we have developed a hybrid predictive framework, integrating LASSO (Least Absolute Shrinkage and Selection Operator) linear regression with decision tree-based modelling. Decision trees are well suited for data analysis when it can be split based on input features, making them a good fit for air pollution analysis, where patterns often vary by time or context (e.g., higher pollution on weekdays compared to weekends). LASSO regression, known for its ability to manage multicollinearity and perform variable selection, assigns weights to coefficients to shrink them toward zero, thereby enhancing model interpretability. In the proposed hybrid model, each node of the decision tree is replaced with a LASSO regression sub-model. Predictions are generated by traversing from the root to the leaf, with final outputs computed through weighted aggregation of model estimates along the path.

While this article focuses on predicting air pollution, LASSO-based predictive models have broad applicability across other critical domains, notably in metallurgy, mining, and mineral processing.[10] As presented in the section 'Discussion and conclusion', these models are well suited for tasks such as geological modelling, process optimisation, quality control, and equipment health monitoring.[11] LASSO modelling in these domains remains a subject for future investigation by the wider research community.

Predictive model design

Node (sensor) configuration

To evaluate the effects of green spaces, vehicular traffic, and human activity on air quality, sensor nodes were strategically deployed across various locations. In 2018, the sensors were installed near the Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje, North Macedonia. Specifically, sensor node $n3'$ was positioned adjacent to a secluded green area, node $n2'$ was located near a vertical green wall structure, and node $n1'$ was placed alongside a high-traffic roadway. Detailed findings on the air quality improvements associated with the green wall installation are presented in Srbinovska et al.[6] In 2019, the sensor layout was modified: node $n1$ was relocated to a pedestrian pavement area, while node $n2$ was positioned near a small green space within the building's internal courtyard. For more information about the position of the sensors and the model of the sensors used for the measurements, as well as their characteristics, see Srbinovska et al.[12] In the same paper, you can find information regarding the data acquisition and preprocessing.

LASSO regression framework

LASSO regression is a linear modelling technique particularly effective in high-dimensional datasets, where the number of predictors is large. It performs both variable selection and regularisation to enhance model interpretability and predictive accuracy. In the context of this study, LASSO is used to model the relationship between meteorological and environmental variables (predictors, independent variables) and air pollution levels, specifically the concentrations of PM2.5, PM10, CO, and NO_2 (target variable).

The LASSO algorithm minimises the residual sum of squares subject to the constraint that the sum of the absolute values of the coefficients is less than a fixed value (often denoted by λ). One of the key features of LASSO is that variable selection is performed by shrinking the coefficients of irrelevant predictors to exactly zero. This means that the most important predictors are automatically selected while the less important ones are disregarded. The amount of regularisation applied is controlled by the tuning parameter λ. A larger λ results in more coefficients being pushed to zero, leading to a simpler model with fewer predictors.

Predictive model: hybrid tree–LASSO approach

The prediction model combines decision tree structures with LASSO regression. The tree is constructed using the CART algorithm[13] as implemented in the *scikit-learn* library.[14] Unlike traditional trees that assign constants at leaf nodes, this hybrid model embeds LASSO linear regressions at every node.

Tree construction:

– Let T be a tree with a root node $t_{0,0}$ and n levels;
– The nodes in level i are defined as follows: $t_{i,j}$ is jth node from left to right;
– Each node contains data that we split into two subsets. For the divided data, one observes 'left child' (one subset of the data) and 'right child' (the second subset of the data). For each subset of the data, one considers linear regression and calculates the mean squared error. From all possible splits of the data set, we are interested in the split that minimises the mean squared error. Hence, one obtains two nodes in the next level, 'left child' and 'right child'. The tree is built until the stopping criteria are satisfied: the tree reaches depth n, or the minimum number of leaves is reached.
– Let t_{n,j_n} be a leaf node (there is no node in level $n+1$ connected to the node t_{n,j_n}). Choose a path from the root node to t_{n,j_n}, $(t_{0,0}, t_{1,j_1}, \ldots t_{n,j_n})$.
– Linear regression model $f_{i,j}$ is fitted to every node $t_{i,j}$ from the tree. For a given λ, the prediction function p is defined as

$$p_{n,j_n} = \frac{f_{0,0} + \lambda f_{1,j_1} + \lambda^2 f_{2,j_2} + \cdots + \lambda^n f_{n,j_n}}{1 + \lambda + \lambda^2 + \cdots + \lambda^n}$$

for the data at the leaf node t_{n,j_n}.

We consider two cases for the parameter λ:

- for small values of λ: the models of the tree that occur higher in the tree are most important, or the models are more global, and

- for bigger λ: the models closer to the leaves become more important, making the model more localised.

Decision tree models typically treat observations as independent and do not inherently consider the temporal continuity present in time series data. To address this, the predictions can be smoothed using standard window-based techniques, such as a Hanning window, with adjustable window sizes to better reflect temporal continuity.

Additionally, the model can be configured to use either the full set of input features or a selected subset. If historical values of the target variable (e.g., past pollution levels) are excluded, the model becomes more suitable for predicting future pollution based solely on external factors, such as weather or traffic. While including historical pollution values can improve short-term prediction accuracy, it also introduces challenges, especially during unexpected events like COVID-19, because such models may accumulate error over time and become unstable when facing sudden shifts or shocks.

Data preparation

In this study, the LASSO regression model is applied to predict the concentrations of PM2.5, PM10, CO, and NO_2, which serve as the target (dependent) variable. Historical averages of these pollutants, specifically 12, 24, 36, 48, 72, 168, 336, and 672 h prior, are incorporated as predictive features. The model is constructed using 20 independent variables, including meteorological factors (temperature, pressure, humidity, wind speed, and wind direction) and environmental indicators (such as workday/weekend and heating/non-heating season). These variables are extracted over various timeframes (1, 24, 48, 72, 168, 336, and 672 h) and are collectively referred to as ‘tree attributes’.

To analyse the potential influence of the COVID-19 pandemic, the dataset is split into two subsets:

Set N: Data collected before 11 March 2020 (pre-COVID, or ‘normal conditions’), and
Set C: Data from 12 March 2020 to 1 March 2022 (during COVID-19 restrictions).
Set N is further divided into training set *T*, which is used for model fitting, and validation set *V*, which is the last 45 days of set *N* and is reserved for tuning and model selection.

Model training and parameter optimisation

In this study, for technical details, see Srbinovska et al.[12] A vast variety of models were developed to predict PM2.5, PM10, CO, and NO_2 concentrations. Models were built under different configurations based on the choice of input attributes and forecasting time intervals. Two attribute sets were considered: a reduced set, which excluded historical values of the target variable, and a full set, which included lagged target values.

The models targeted different time intervals, specifically: 1 hour (1h), 1 day (1d), 2 days (2d), 3 days (3d), 1 week (1w), and 2 weeks (2w). For each configuration, models were trained on a designated training set *T*, and a comprehensive grid search was performed over several parameters. The best-performing model, based on validation performance, is selected for final evaluation on the test set *C*.

The parameters included:

- Parameter λ varied in the range [0.2, 3.0], controlling the influence of local vs. global linear models;
- Smoothing window size varied from 1 to 168 (hours), corresponding to smoothing periods ranging from 1 h to 3 weeks;
- Two smoothing techniques were evaluated: uniform (ones) and Hanning window[13];
- Maximum tree depth, varied from 0 to 8, controlling tree complexity;
- Minimum number of samples per leaf, set between 80 and 150, to ensure model stability and prevent overfitting.

Each node in the decision tree included a linear regression model, without restriction on the sign of the coefficients, allowing both positive and negative influences of predictors. As a result, for each combination of time interval and attribute set, a single best-performing model was selected for evaluation on the final test set *C*.

The primary objective of this study is to determine the most effective predictive model for PM2.5, PM10, CO, and NO_2 pollution levels in the city of Skopje and subsequently to evaluate whether the COVID-19 pandemic had a measurable impact on air quality. To investigate the second objective, the best-performing validated models, identified through training on dataset *T*, were applied to the post-pandemic period data contained in dataset *C*.

To assess deviations, the difference between the predicted values $\hat{y}$ and the actual recorded pollution values y in C was calculated. This discrepancy is denoted as PmR (pollution model residual). A positive PmR suggests that the recorded pollution levels were lower than expected, indicating a positive impact of the COVID-19 pandemic on air quality. Conversely, a negative PmR suggests a worsening of air quality relative to predictions, while a PmR value close to zero implies no substantial deviation.

To benchmark model performance, the relative error (RE) metric is used. RE is computed as the ratio of the sum of squared residuals of the model to the sum of squared residuals from a trivial model that predicts the mean of the target variable. This metric provides a normalised view of performance relative to a baseline. Lower RE values indicate stronger predictive performance, with $RE = 0$ representing a perfect model and $RE = 1$ meaning the model performs no better than the mean predictor.

Results

This paper aims to identify the most effective prediction model for PM pollution in the city of Skopje and to evaluate whether air quality was affected by the COVID-19 pandemic. To address the second objective, the validated models are applied to the corresponding data in Set C. For the technical methodology underlying this study, we refer the reader to a recent publication by the same authors.[12]

In the same paper, the results obtained from the selected validated models for PM2.5 and PM10 are given. Two

Table 1. Results for CO and NO_2 for a full set of attributes and a reduced set of attributes.

	Full set of attributes				Reduced set of attributes			
	CO		NO_2		CO		NO_2	
Target	RE	PmR	RE	PmR	RE	PmR	RE	PmR
$n1-1h$	0.06	0.24	0.07	0.48	0.11	0.09	0.04	−0.53
$n1-1d$	0.03	0.08	0.05	0.0	0.1	0.02	0.07	−0.16
$n1-2d$	0.03	0.06	0.05	0.14	0.09	0.01	0.1	−0.28
$n1-3d$	0.04	−0.1	0.02	0.04	0.08	0.03	0.05	−0.22
$n1-1w$	0.16	0.62	0.02	0.14	0.02	0.01	0.06	−0.32
$n1-2w$	0.07	−0.32	0.0	0.05	0.01	0.04	0.02	−0.26
$n1-4w$	0.01	−0.05	0.03	0.02	0.01	−0.46	0.16	−0.16
$n2-1h$	0.45	−0.06	0.09	0.56	2.66	−0.1	0.15	−1.04
$n2-1d$	0.22	−0.04	0.04	0.24	1.09	−0.17	0.11	−0.69
$n2-2d$	0.27	−0.03	0.04	0.32	3.56	−0.32	0.02	−0.84
$n2-3d$	0.45	−0.04	0.04	0.1	4.77	−0.07	0.02	−0.78
$n2-1w$	1.89	−0.14	0.04	−0.02	4.68	0.11	0.02	−0.84
$n2-2w$	1.05	−0.1	0.02	0.32	4.00	−0.1	0.03	−0.77
$n2-4w$	3.11	−0.07	0.01	0.44	1.63	−0.08	0.0	−0.76
$n3-1h$	0.1	−0.05	0.42	0.54	0.58	−1.31	0.08	−0.8
$n3-1d$	0.06	−0.21	0.05	0.62	0.45	−1.49	0.06	−0.46
$n3-2d$	0.04	−0.08	0.09	0.43	0.5	−1.53	0.03	0.43
$n3-3d$	0.06	−0.12	0.16	0.07	0.62	−1.09	0.01	−0.38
$n3-1w$	0.02	0.00	0.08	0.81	0.18	−1.1	0.33	−1.09
$n3-2w$	0.03	−0.15	0.12	0.72	0.06	−1.14	0.04	−1.19
$n3-4w$	0.02	−0.19	0.04	0.09	0.01	−1.31	0.01	−1.14

Note: The 'Target' column indicates the sensor node's location and the time frame used in the analysis. PmR stands for pollution model residual, while RE is the relative error metric.

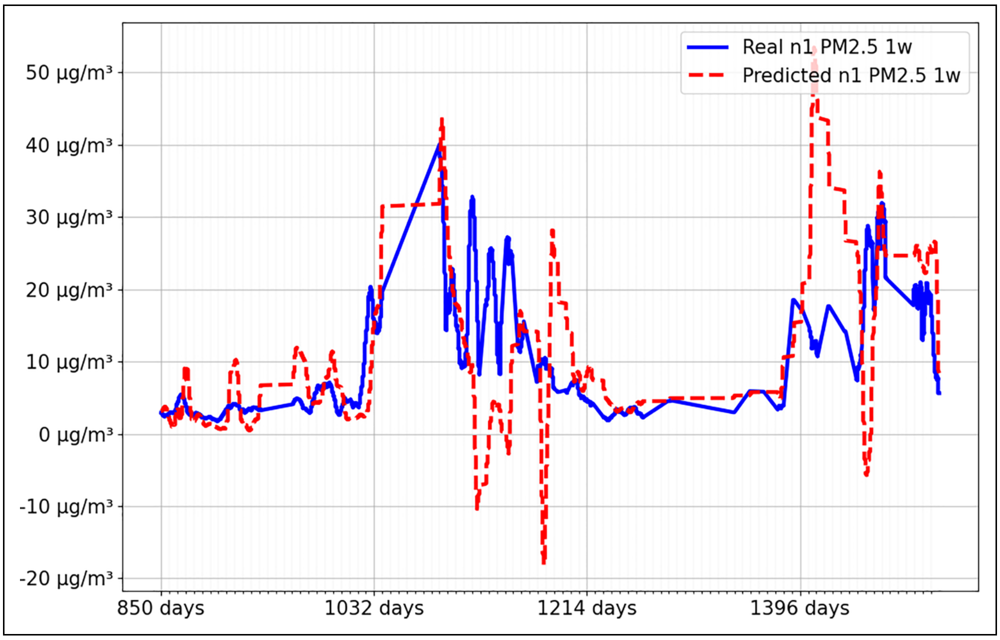

Figure 1. PM2.5-predicted values (dashed line), real values (continuous line), using historical values.

different approaches are considered: one using the full set of attributes, which includes historical data from the recent past, and another employing a reduced set of attributes, where predictions are made without historical values. In Table 1, the results for the pollutants CO and NO_2 are presented.

We focus on a clearer visualisation of the obtained results by presenting comparative charts of the recorded versus predicted values using the optimal models, specifically for data collected (for PM2.5, PM10, CO, and NO_2) from sensor node $n1$.

Figure 1 presents the predicted (dashed red line) vs. real (continuous blue line), that is, recorded values of PM2.5 for a target interval of 1 week. The model uses historical values. The best model has parameters λ 3, and the maximum depth of the tree was set to 1. The predictions were smoothed with a window of size 36.

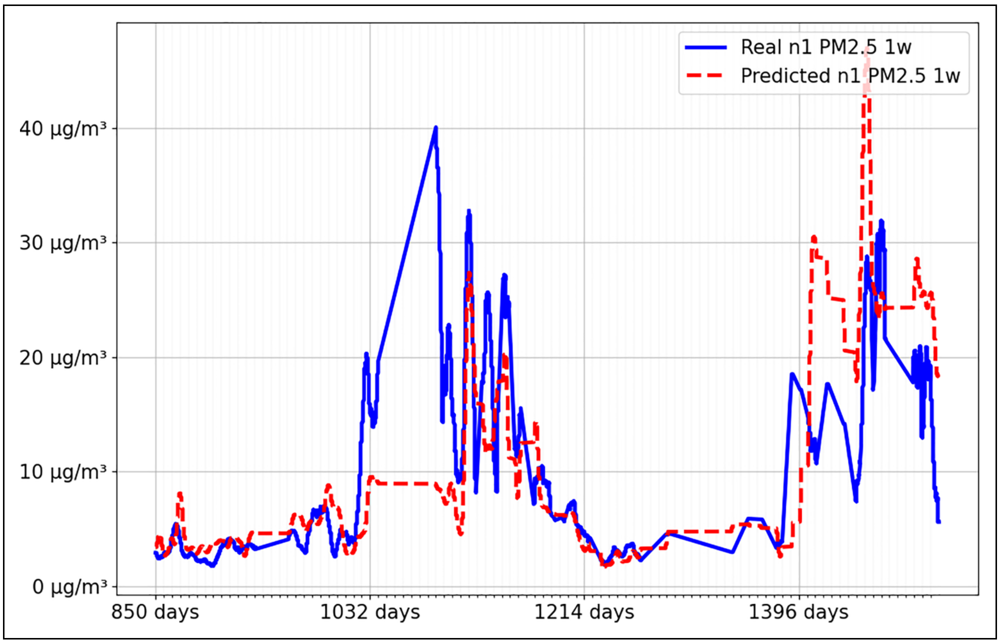

Figure 2. PM2.5-predicted values (dashed line), real values (continuous line), without historical values.

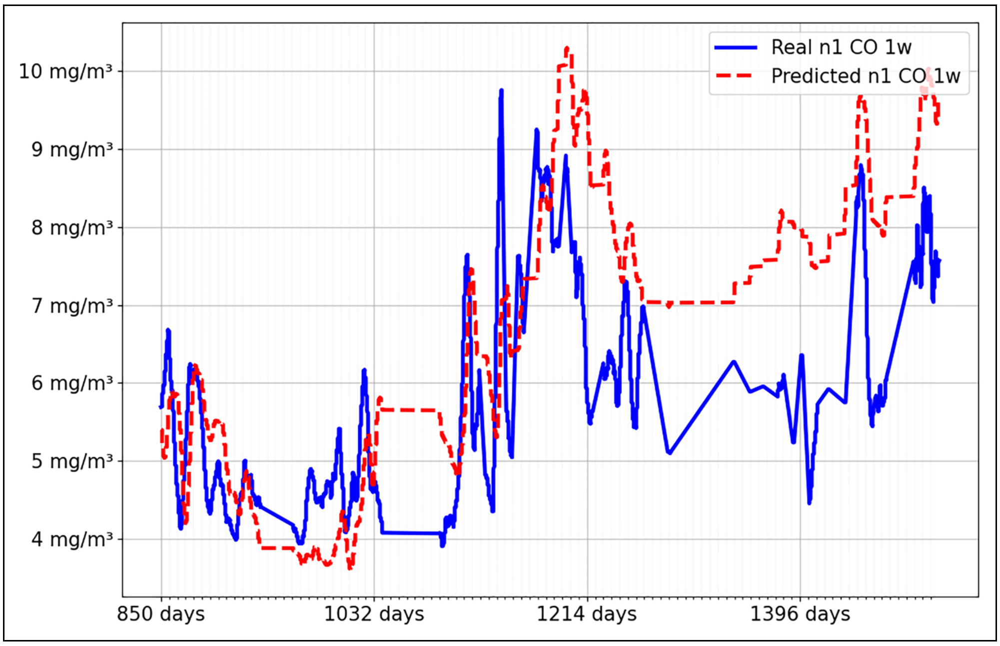

Figure 3. CO-predicted values (dashed line), real values (continuous line), using historical values.

In Figure 2, the predicted vs. real values of PM2.5 for a target interval of 1 week and no historical data are presented. The best model has parameters λ 2, maximum depth of the tree set to 5, and minimum number of samples per leaf set to 120. The predictions were smoothed with a *Hanning* window of size 168. Figures 3 and 4 present the results for CO, with and without historical data, respectively. Figure 3 presents the predicted vs. real values of CO for a target interval of 1 week, using a model that incorporates historical data. The best-performing model in this case is characterised by $\lambda = 0.2$ and a tree depth of 1. The model's coefficients can be both positive and negative. To enhance smoothness, the predictions are averaged using a window length of 168 with the Hanning method, which applies a standard mean.

Figure 4 presents the predicted vs. real values of CO for a target interval of 1 week, using a model without historical data. The best-performing model in this case was achieved with $\lambda = 0.5$, a maximum tree depth of 1, and a minimum of 150 samples per leaf. The linear models were constrained to have only positive coefficients. Finally, the predicted values were smoothed using a Hanning window with a length of 168. Figure 5 presents the predicted vs. real values of PM10 for a target interval of 1 week, using a model with historical data. The best-performing model in this case was achieved with $\lambda = 1.0$, a maximum tree depth of 4, and a minimum of 80 samples per leaf. The predicted values were smoothed using a Hanning window with a length of 168.

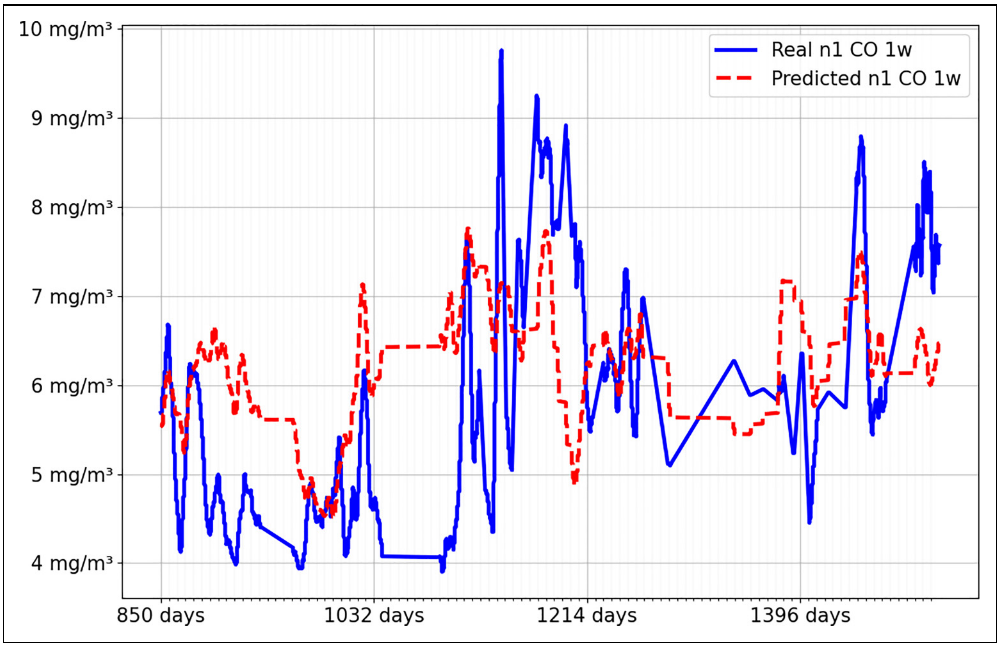

Figure 4. CO-predicted values (dashed line), real values (continuous line), without historical values.

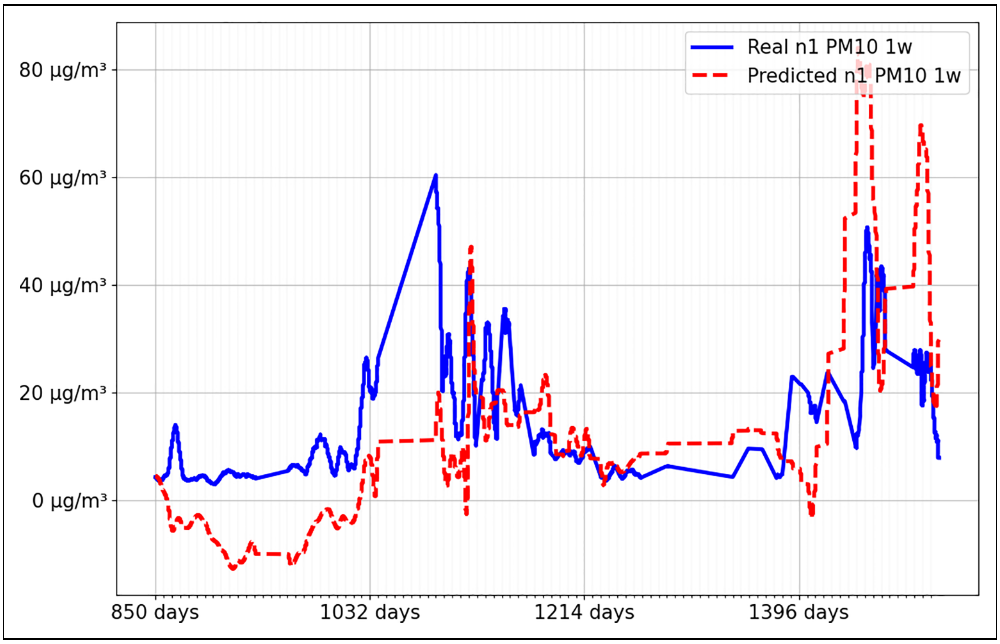

Figure 5. PM10-predicted values (dashed line), real values (continuous line), using historical values.

Figure 6 presents the predicted vs. real values of PM10 for a target interval of 1 week, using a model without historical data. The best-performing model in this case was achieved with $\lambda = 1.5$, a maximum tree depth of 8, and a minimum of 150 samples per leaf. The predicted values were smoothed using a Hanning window with a length of 168.

Figure 7 presents the predicted vs. real values of NO_2 for a target interval of 1 week, using a model with historical data. The best-performing model in this case was achieved with $\lambda = 2.0$, a maximum tree depth of 4, and a minimum of 150 samples per leaf. The predicted values were smoothed using a Hanning window with a length of 168.

Figure 8 presents the predicted vs. real values of NO_2 for a target interval of 1 week, using a model without historical data. The best-performing model in this case was achieved with $\lambda = 3.0$, a maximum tree depth of 8, and a minimum of 150 samples per leaf. The predicted values were smoothed using a Hanning window with a length of 168.

Other model comparison. We experimented with artificial neural networks (ANNs), but the results consistently suffered from overfitting. In contrast, our hybrid approach generally avoids this issue by relying on an ensemble of smaller regression models. Our objective was to evaluate model predictions in a 'what-if' scenario, which guided our model selection.

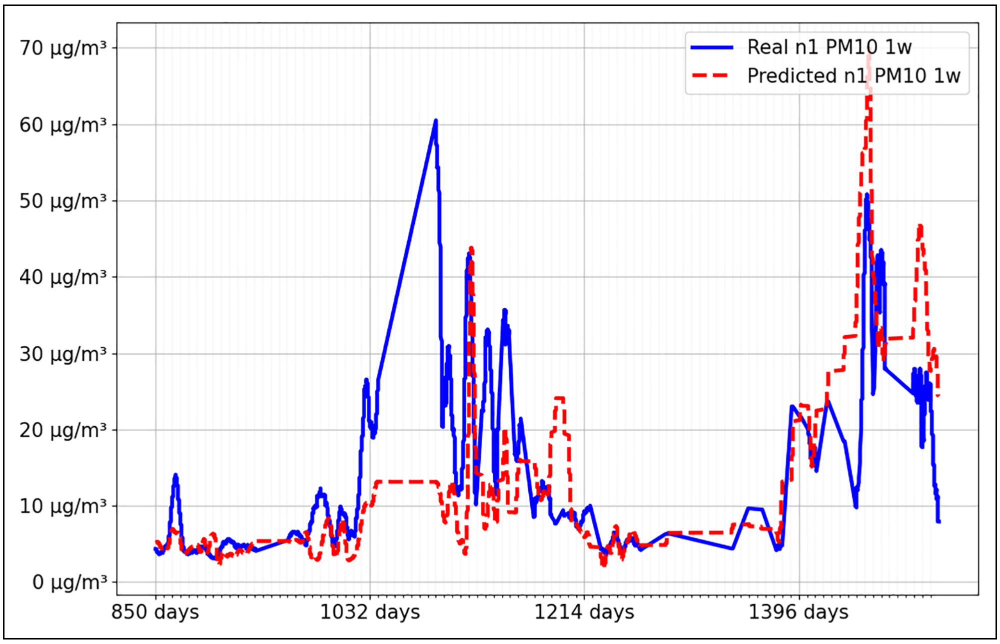

Figure 6. PM10-predicted values (dashed line), real values (continuous line), without historical values.

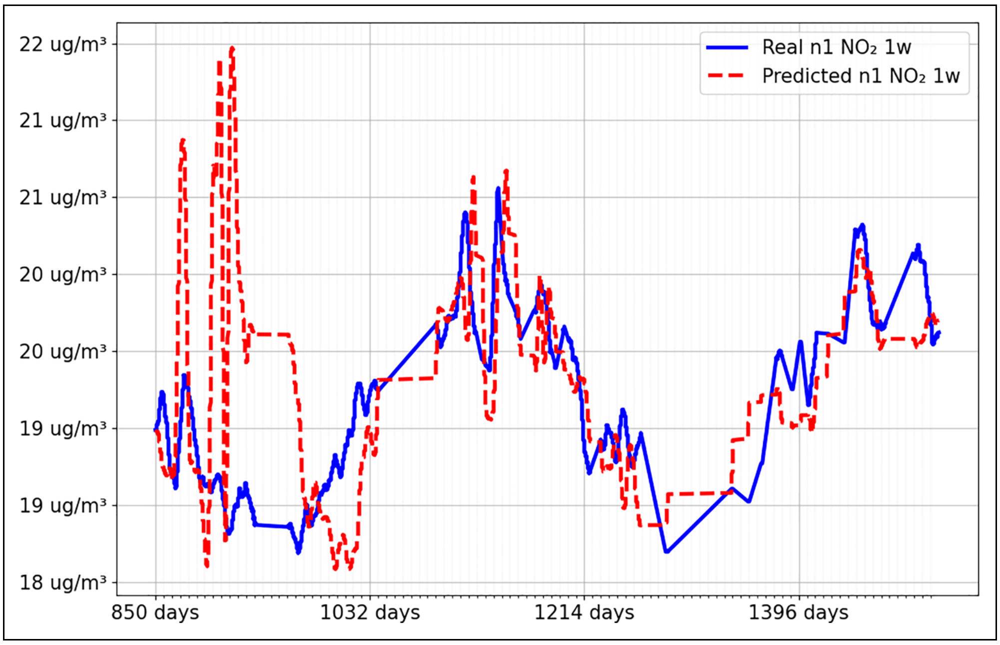

Figure 7. NO_2-predicted values (dashed line), real values (continuous line), using historical values.

Simple linear regression would have been too limited, while neural networks, in the ANN case, tended to overfit the data. That said, we did not explore recurrent architectures such as RNNs or LSTMs. Exploring these alternatives could be valuable in future work, particularly in a different research context.

Discussion and conclusion

Our special interest lies in the applications of our hybrid tree-LASSO model within the context of mineral processing and extractive metallurgy, where complex, multivariate data and sensor-driven environments are common. Our model, exploited in the sections 'Predictive model design' and 'Results', can be analogously applied across key unit operations in mineral processing. Just as in pollution modelling, mineral processing and extractive metallurgy involve highly nonlinear, multivariate processes influenced by a wide range of operational and material variables. We have reasonable expectations (our model is a generalisation of the classical model) that the model will not only replicate classical LASSO performance but also extend it, offering more nuanced insights and better adaptability to complex industrial data.

Consider flotation, where performance hinges on a delicate balance of variables like pH, reagent dosage, and particle size. Our model cuts through the noise to highlight

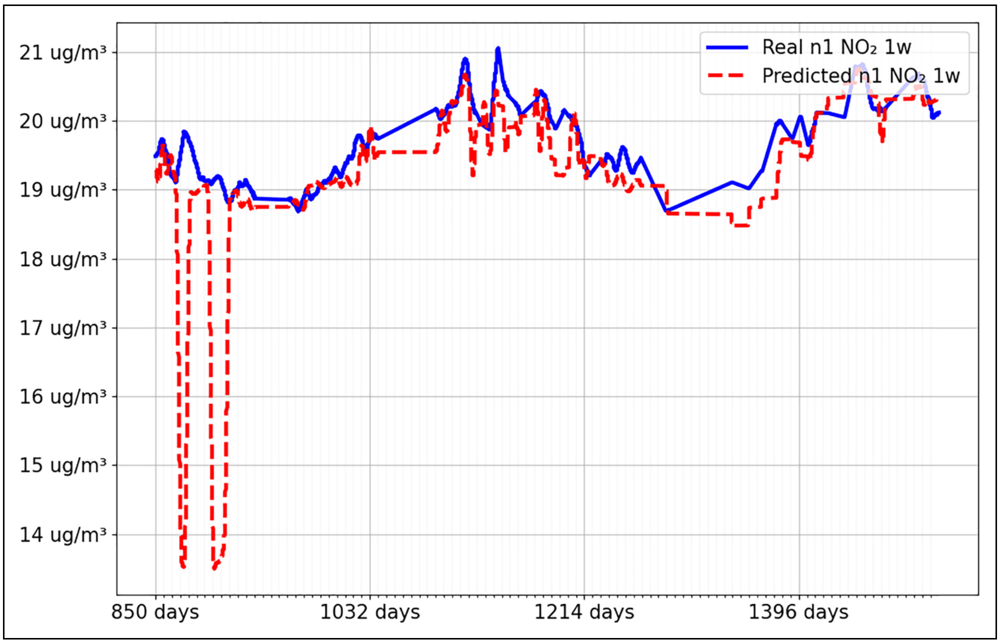

Figure 8. NO_2-predicted values (dashed line), real values (continuous line), without historical values.

the most influential factors, improving both recovery rates and process stability. With the noise stripped away, operators gain a clearer view of what matters most, leading to steadier process control and better recovery outcomes. A similar advantage appears in grinding and milling circuits, where the model helps identify the parameters that truly affect throughput and particle size, supporting energy-efficient and longer-lasting operations.

In hydrometallurgical leaching, where chemical and mineralogical complexity often clouds analysis, the model isolates predictive variables that optimise reagent use and reduce impurities. Likewise, in high-dimensional spectroscopic data from technologies like laser induced breakdown spectroscopy, near infrared analysis, and X-ray fluorescence, the model selects the most informative wavelengths, enabling rapid and accurate predictions for sorting or blending.

Even in environmentally sensitive operations like heap leaching, the model supports reliable predictions of recovery and effluent quality. Finally, in equipment maintenance, it pinpoints early indicators of failure like vibration, heat, and motor current, allowing for proactive, data-driven upkeep.

We finally conclude the article with the main results of this study. A hybrid predictive modelling approach combining decision trees and LASSO-regularised linear regression was applied to air pollution forecasting. The use of LASSO allowed for automatic feature selection by penalising large coefficients, effectively reducing model complexity and mitigating overfitting capabilities not offered by standard linear regression.

The dataset used in the analysis comprises air quality measurements collected from three sensor nodes located on the technical campus of Ss. Cyril and Methodius University in Skopje, covering the period from 2018 to 2022. The onset of the COVID-19 pandemic in March 2020 introduced substantial changes in human activity, including reduced traffic and the closure of schools, universities, industrial facilities, and public spaces. To assess the impact of these changes on air quality, a wide range of predictive models were trained and validated using only pre-pandemic data, that is, data collected before the enforcement of COVID-19-related restrictions in the Republic of Macedonia.

The modelling framework explored various parameters, including multiple values of the regularisation parameter λ, different decision tree depths, two smoothing techniques (uniform and Hanning windows), and multiple prediction time frames. In total, 2520 distinct models were generated and evaluated for each target variable. Using the best-performing models, the concentrations of PM2.5, PM10, CO, and NO_2 were predicted based on meteorological and environmental attributes, both with and without incorporating historical pollutant concentration values. The predicted values were then compared to actual recorded measurements during the post-COVID period. Results showed that models using the full feature set, including historical concentration data, achieved higher predictive accuracy. However, in both modelling scenarios, the predicted pollution levels were consistently higher than the observed values during the pandemic period. This systematic overestimation suggests a positive impact of the COVID-19 pandemic on air quality in Skopje, likely attributable to the reduced anthropogenic activity.

The developed models demonstrate the capability to predict pollution levels in the post-COVID period and can be readily extended to other pollutants and additional geographic locations, making them versatile tools for environmental monitoring and public policy evaluation.

ORCID iDs

Mare Srbinovska https://orcid.org/0000-0003-4514-8562
Pavel Dimovski https://orcid.org/0000-0003-0715-9498
Maja Celeska Krstevska https://orcid.org/0000-0003-4019-5378
Vesna Andova https://orcid.org/0000-0001-9455-1345
Sijche Pechkova https://orcid.org/0000-0001-6122-6482

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Samet J, Zeger S, Dominici F, et al. The national morbidity, mortality, and air pollution study. Part II: morbidity and mortality from air pollution in the United States. *Res Rep Health Eff Inst* 2000; 94: 5–79.
2. Meisner C, Gjorgjev D and Tozija F. Estimating health impacts and economic costs of air pollution in the Republic of Macedonia. *South East Eur J Public Health* 2015; 4.
3. European Environment Agency. Europe's air quality status 2024, https://www.eea.europa.eu//publications/europes-air-quality-status-2024 (accessed 30 June 2025).
4. Jacobson MZ. Effects of meteorology on air pollution. In: Jakobson MZ (ed) *Atmospheric pollution: history, science, and regulation*. Cambridge: Cambridge University Press, 2002, pp.145–178.
5. Alberto A, Ramos N and Almeida R. Parametric study of double-skin facades performance in mild climate countries. *J Build Eng* 2017; 12: 87–98.
6. Srbinovska M, Andova V, Krkoleva Mateska A, et al. The effect of small green walls on reduction of particulate matter concentration in open areas. *J Cleaner Prod* 2021; 279: 123306.
7. Ju MJ, Oh J and Choi YH. Changes in air pollution levels after COVID-19 outbreak in Korea. *Sci Total Environ* 2021; 750: 141521.
8. Wu CL, Wang HW, Cai WJ, et al. Impact of the COVID-19 lockdown on roadside traffic-related air pollution in Shanghai, China. *Build Environ* 2021; 194: 107718.
9. Dimovska M and Gjorgjev D. The effects of COVID-19 lockdown on air quality in Macedonia. *Open Access Maced J Med Sci* 2020; 8: 353–362.
10. Kang Y, Ren MM, Zhao JX, et al. Prediction of end-point phosphorus content of molten steel in BOF with machine learning models. *J Min Metall Sect B* 2024; 60: 93–103.
11. Guan PW, Debusschere BJ, Bishop SR, et al. LASSO For CALPHAD model selection enables data-efficient thermodynamic modeling: an application in thermochemical hydrogen production materials. *ACS Appl Energy Mater* 2025; 8: 8589–8597.
12. Srbinovska M, Pechkova S, Pechkov A, et al. A comprehensive study on air pollution prediction in North Macedonia: insights from LASSO modeling. IEEE Explore. 2024. https://ieeexplore.ieee.org/document/10645136.
13. Berk RA. Classification and regression trees (CART). In: Berk RA (ed) *Statistical learning from a regression perspective*. Cham: Springer Texts in Statistics. Springer, 2016, pp.129–186.
14. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011; 12: 2825–2830.