JOURNAL OF ENGINEERING MANAGEMENT AND COMPETITIVENESS (JEMC)
VOL. 11, NO. 2, 2021, 115-123

INTERPRETING RELATIONSHIPS BETWEEN POLLUTANTS AND CARBON DIOXIDE EMITTED INTO AIR FROM INDUSTRIES IN SERBIA

UDC: 502.3:504.5]:519.233.5
Original Scientific Paper

Bulent TUTMEZ[1]

[1]Inonu University, Faculty of Engineering, 44280 Malatya, Turkey
E-mail: bulent.tutmez@inonu.edu.tr

Paper received: 09.07.2021.; Paper accepted: 10.09.2021.

The focus was on the pollution problem in Serbia and the relationships between CO_2 emitted into air from industries and air quality indicators such as particulate matters (PM2.5, PM10), nitrogen and sulfur oxides (NO_x, SO_x), and volatile organic compounds were analyzed. To identify the dependencies, both parametric and nonparametric statistical learning-based evaluation algorithms were taken into consideration. Both the model structures produced satisfactory estimations with high accuracy levels. As a result of the model interpretation, PM2.5 has been recorded as the main indicator to explore the variability in CO_2 concentrations. The implementations exhibited that interpretable machine learning can provide meta-data and sufficient information for making black-box air quality system more explainable. Thus, the practiced modelling tools, the provided interrelationships as well as the new information could be considered by the national authorities within a computational environmental management strategy.

Keywords: Greenhouse gas; Air pollution; Statistical learning; Regression; Variable importance.

INTRODUCTION

Air pollution can influence health and it has many long and short-term effects such as cancer, cardiovascular diseases, and respiratory diseases. Recently, strong relationship between air pollution and Covid-19 pandemic has been underlined (Brauer et al. 2021). Providing public air quality is one of the most important environmental management strategies in public management. Managing air quality covers control strategies to perform air pollution reduction. To make them actual the pollution prevention approaches, determination and control of the amounts of emissions released by industrial production and consumptions have critical importance. Ultimately, air quality and climate change are functionally linked from their emission sources to their effects on human health and ecosystems (Melamed, Schmale, & Schneidermesser, 2016).

The number of national and global agreements to focus on environmental conversion and sustainability issues is tending to increase (Liu, & Shen, 2014). There is also a strong relationship between the policies and air pollution measures as well as GHG emissions. According to an OECD report (Organization for Economic Cooperation and Development) published in 2012, air pollution would be the top environmental reason of mortality worldwide by 2050 (OECD, 2012). Therefore, both the national authorities and the international organizations deal more with potential impacts and necessary precautions in recent years.

The greenhouse gas (GHG) effect addresses the natural warming of the earth that ends up when gases in the atmosphere trap heat arising from the sun. GHG emissions such as carbon dioxide (CO_2), methane (CH_4) and nitrous oxide (N_2O) hold the surface temperature of the Earth. The GHGs affect global climate conditions and without this impact, the Earth would be a frozen globe. Among the GHGs, CO_2 is the most plentiful pollutant that contributes to climate change by use of fossil fuel combustion (Manisalidis et al., 2020; S. Paraschiv, & L. S. Paraschiv, 2020). Similar to GHGs, various air pollution indicators based upon

ISSN 2217-8147 (Online)

Available online at http://www.tfzr.uns.ac.rs/jemc

industrial production and consumptions can be identified such as particulate matters (PM2.5, PM10), nitrogen and sulfur oxides (NO_x, SO_x), ozone, carbon monoxide (CO) as well as volatile organic compounds. The main industrial sources of these pollutants are fuel combustion sources, refineries, cement kilns, quarries and transportation components like highway vehicles (Cristescu, Stoica, & Suditu, 2019).

Even though many investigation outputs were presented in literature on greenhouse gases and air pollution indicators, there are limited studies that handled them together. One of the prominent studies, Zhang et al. (2013) focused on PM10 and CO_2 emissions in Tianjin, China based on electricity industry. In an analytical policy-based study, Bollen and Brink (2014) suggested an analytical framework for air pollution control in Europe by combining greenhouse gases and air pollutants. Recently, an emission reduction analysis for freight transportation and containers has been presented based on the parameters: CO_2, NO_x and PM10 (Bal, & Vleugel, 2020).

Beyond this general scientific literature, the use of statistical techniques and machine learning (ML) algorithms in air pollution control and GHGs has gained popularity in recent years. One of the antecedent studies, Ni et al. (2017) evaluated agricultural air quality in Indiana, USA by national air emissions and statistical analyses. In this study, both CO_2 and some pollutants (NH_3, H_2S) were considered. Bellinger et al. (2017) reviewed the ML techniques used in air pollution problem and epidemiology. Air pollution emissions recorded in Spain were predicted by random forest and hierarchical clustering (Martinez-Espana et al., 2018). Similarly, air quality in urban areas of Bulgaria was evaluated by ARIMA and random forest models (Gocheva-Ilieva, Ivanov, & Livieris, 2020). Recently, the air quality assessment of Eskisehir city in Turkey has been performed by statistical learning-based regularization (Tutmez, 2020).

At this stage, understanding the relationships between air pollutants and CO_2 could be critical both at regional and global scales. Due to heterogeneous nature of air pollution environment and multi-source data properties, this modelling attempt requires a holistic perspective including model accuracy and interpretation in together. For this purpose, supervised regression analyses have been identified as optimal to explore the dependencies from linear and non-linear approaches. Exploring the relationships between CO_2 emitted into air from industries and air quality indicators is still a novel problem type and there is no ML-based study in literature focusing on this identification problem. Therefore, dependencies between the major greenhouse gas (CO_2) and the major air quality measures have been handled from a statistical machine learning framework. To make realistic and multi-directional analyses, both linear and non-linear regression algorithms have been utilized. Both model types provided high accuracy and meta-data. The numerical outcomes revealed that relatively small particulate matter and nitrogen oxide are the main indicators for understanding the variability in carbon dioxide concentrations resourced from different industrial sources and consumptions.

The rest of the article is structured as follows. Section 2 introduces the problem and the methodological ground used in this study. Section 3 gives the numerical experiments; the results and a brief discussion are given in this section. Section 4 summarizes the findings of the investigation.

METHODOLOGY

Parametric analysis: principal component regression (PCR)

In a multivariate regression analysis, use of limited number of variables in a model is necessary for simplicity. PCR decreases the number of indicator variables and solves data collinearity problem. This approach decomposes a data matrix X into scores T and loadings P as follows (Varmuza, & Filzmoser, 2009):

$$X = TP^T + E\,. \tag{1}$$

The score matrix T includes the maximum amount of information of data matrix based on orthogonal projections on linear combinations. In Eq. (1), E denotes a residual matrix. The structure can be clarified by the conventional regression form:

$$y = Xb + e\,. \tag{2}$$

In PCR analysis, the matrix X in Eq. (2) is replaced by T. In this way, main information of the x-data for regression on y is represented (Suryanarayana,

& Mistry, 2016). The new regression structure and coefficients can be obtained by:

$$y = Tg + e_T, \quad (3)$$

$$g = P^T b. \quad (4)$$

The structure given in Eq. (4) has ability to remove collinearity and produce new score vectors. Thus, the PCR model coefficients can be provided as follows:

$$b_{PCR} = Pg. \quad (5)$$

Using the PCR approach in air quality modelling can provide some advantages such as solving multicollinearity problem and a notable improvement in testing predictive accuracy compared with the conventional multiple linear regression (MLR) approach. It should also be noted that the PCR modelling simply seeks to decrease the variability present throughout the predictor space.

Nonparametric Analysis: multivariate adaptive regression splines (MARS)

MARS addresses a multivariate-additive model and nonparametric flexible solution. It focuses on approximation of the relationship between independent variables, *x* and response variable, *y* using piecewise linear basis functions (BFs) (Hastie, Tibshirani, & Friedman, 2017):

$$(x-t)_+ = \begin{cases} x-t, & if\ x > t, \\ 0, & otherwise, \end{cases}$$

$$\text{and } (t-x)_+ = \begin{cases} t-x, & if\ x < t, \\ 0, & otherwise. \end{cases} \quad (6)$$

where, *t* denotes the critical point as with a knot. The "+" identifies positive part.

A MARS algorithm follows a two-step implementation consisting of forward selection and backward deletion procedures. In the forward process, the algorithm determines BFs that are added to the regression model by a fast searching. This process comes to rest when the algorithm provides the maximum number of BFs. In the backward process, the model is pruned to reduce the overall performance and the BFs are detracted from the model. For this operation, the algorithm utilizes the generalized cross-validation (GCV). Thus, an optimally estimated regression structure is provided (Özmen, 2016). The following functional regression structure gives the model (Zhang, & Goh, 2016):

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(x), \quad (7)$$

where, $\beta_m B_m(x)$ represents a BF, and *M* denotes the number of BFs in the regression model.

Use of the MARS approach in air modelling analysis also includes a potential for provision of some reliable outcomes. Different from other nonlinear models, it obtains an intuitive stepping block into nonlinearity after grasping the structure of the MLR (Boehmke, & Greenwell, 2020).

CASE STUDY

Study Data

Health impact of ambient air pollution in Serbia has been reported by the World Health Organization. According to the report, an estimate of 6 592 deaths and 131 183 years of life lost (YLL) were due to air pollution in Serbia (WHO, 2019). Appraisals of air quality based on measurements derived from monitoring stations are managed by the Serbian national authority. The Statistical Office of the Republic of Serbia takes an active role in data collection and management together with other institutes such as the *Dr. Milan Jovanovic Batut* Institute of Public Health of Serbia and the Serbian Environmental Protection Agency (SEPA). The application data utilized in modelling works were provided from the Statistical Office Database (SORS, 2020). The data set includes air pollution and greenhouse gas emissions measured in 2014 sourced from 62 different main industries in Serbia. The data published broken down by NACE (the industry standard classification system used in the European Union) classification of economic activities. The data set addresses the calendar year. The air emissions accounts are presented in tonnes as Mg (milligram).

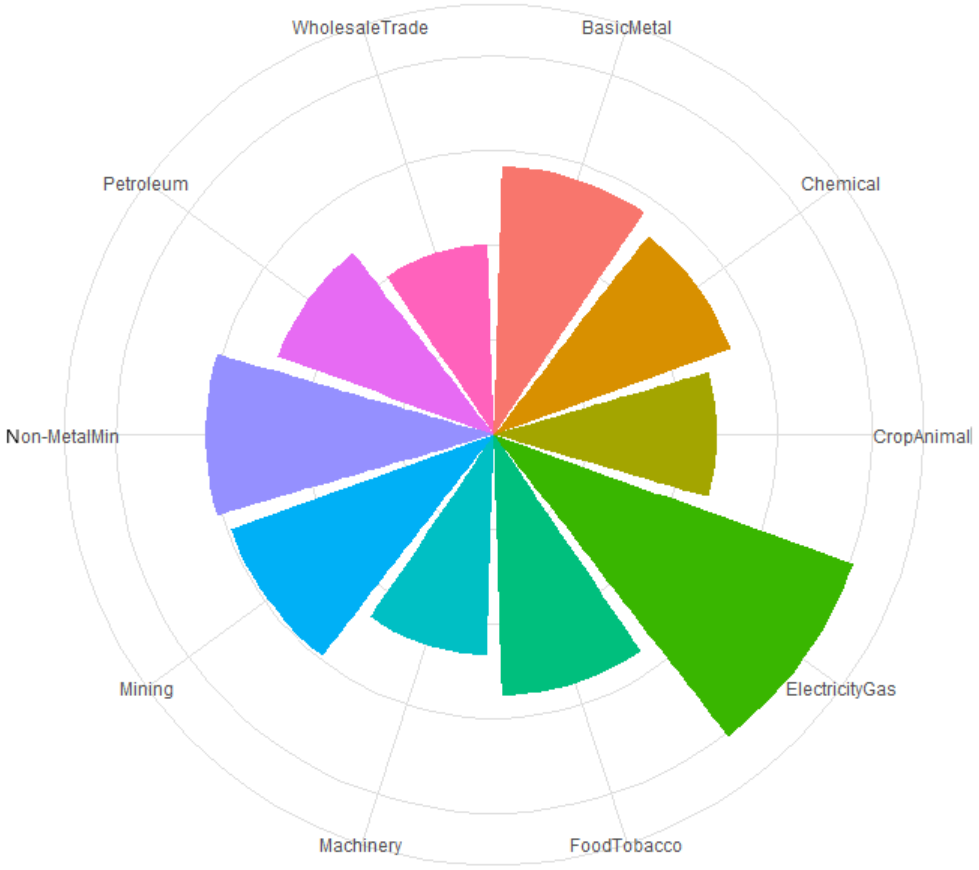

Figure 1: Main pollutant industries in Serbia

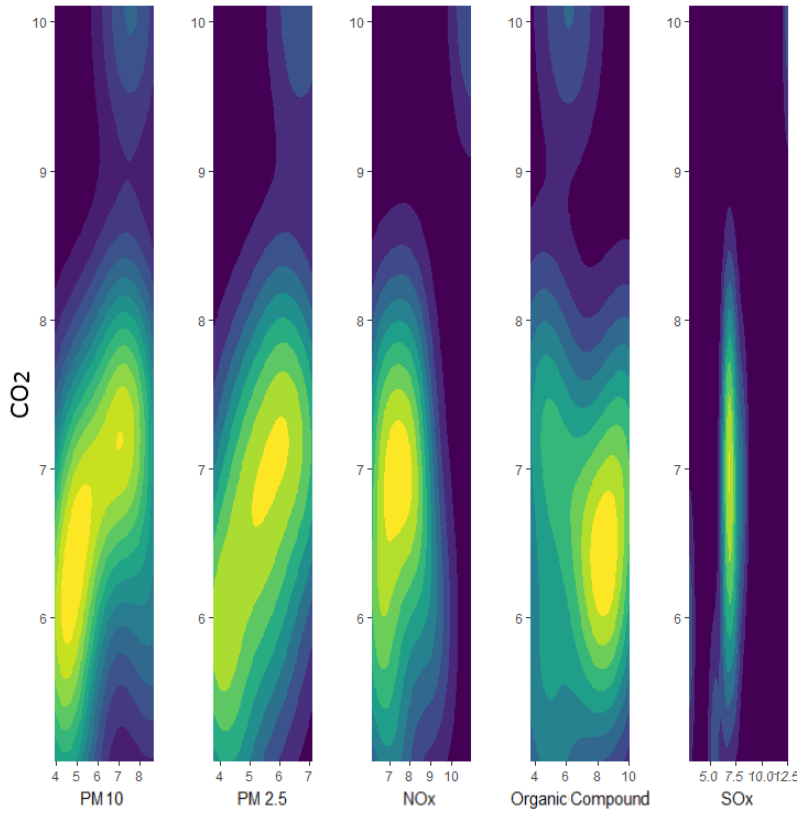

Figure 2: Contour plots for air quality indicators and CO_2

Structure Identification

The industrial pollution sources in Serbia are categorized by the national authority. Because the industry-specific sources have different level impacts, a logarithmic transformation has been performed. Figure 1. developed by the application data shows the impact levels of the main sectors on air pollution heavily. In addition to electricity-gas-air conditioning category, non-metallic manufacturing and mining-quarrying come to the forefront industries in Serbia.

To explore the relationships between CO_2 and air quality characteristics (PM10, PM2.5, NO_x, Organic Compound and SO_x) the measured projections have been exhibited first using contour maps in Figure 2. The domains around average CO_2 concentrations have similar levels and shapes. Among the parameters, SO_x produces relatively different structure. It could partly be due to scale of the horizontal axis. In addition, burning of the fossil fuels is the major sources of this toxic gas and a steady emission could be expected.

Results and Discussion

PCR-based parametric analysis was carried out following a two-step procedure. First, uncorrelated components were provided. After that, the resulting features were employed as estimators in linear model. The 10-fold cross validation (CV) on the PCR model revealed the optimal number of components based on Root Mean Squared Error measures (RMSE) (Figure 3). Although the optimization process addresses one feature, the use of two or three components having relatively low error levels could produce satisfactory results as well.

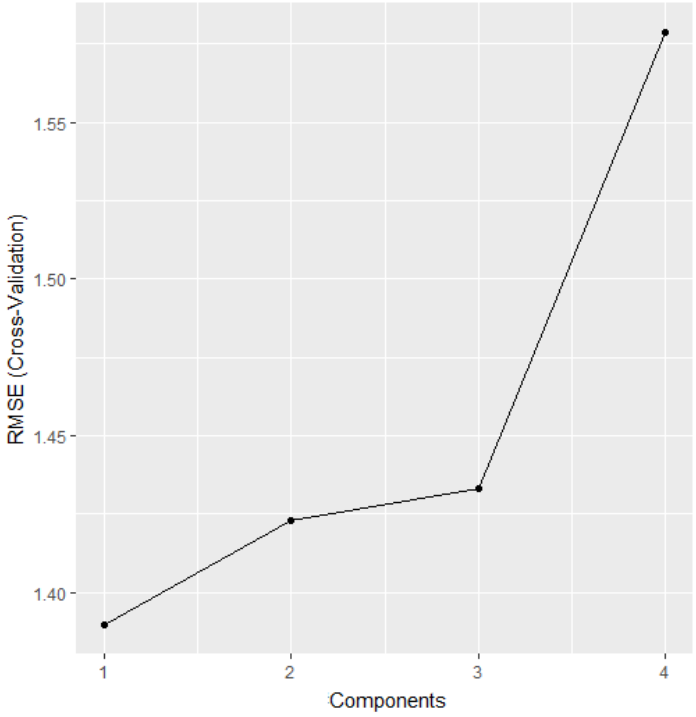

Figure 3: 10-fold CV RMSE provided by PCR

The nonlinear regression analyses were carried out using spline-based adaptive regression. Before the CV-optimized structure, an implementation was conducted and variation of the residuals was recorded. In this way, the potential outliers and an overview of the performance of the MARS model have been checked. Figure 4 displays the residuals at different levels using a cumulative distribution plot. In a grid search by 10-fold CV, the focus was on interaction of the first and the second degrees. The Generalized Cross Validation (GCV) approach has been considered during the optimization process. The MARS model that obtains the optimal combination consists of the first-degree interaction effects and retains 4 terms (Figure 5.).

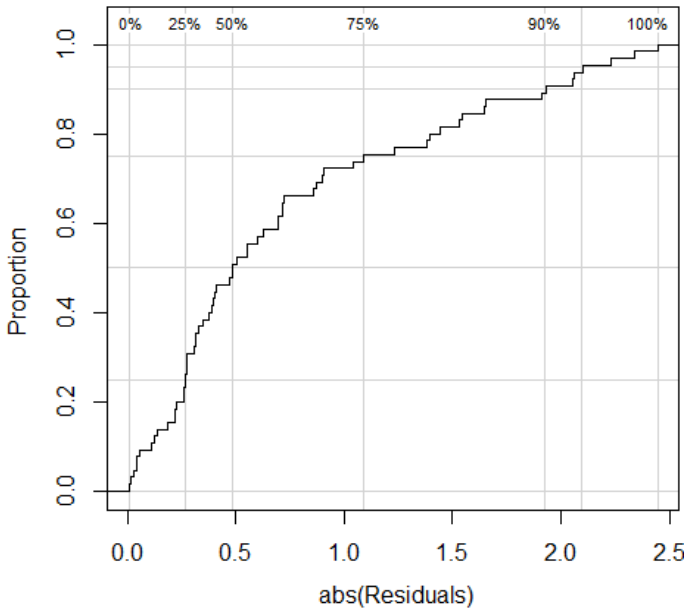

Figure 4: Cumulative distribution plot for residuals of MARS model

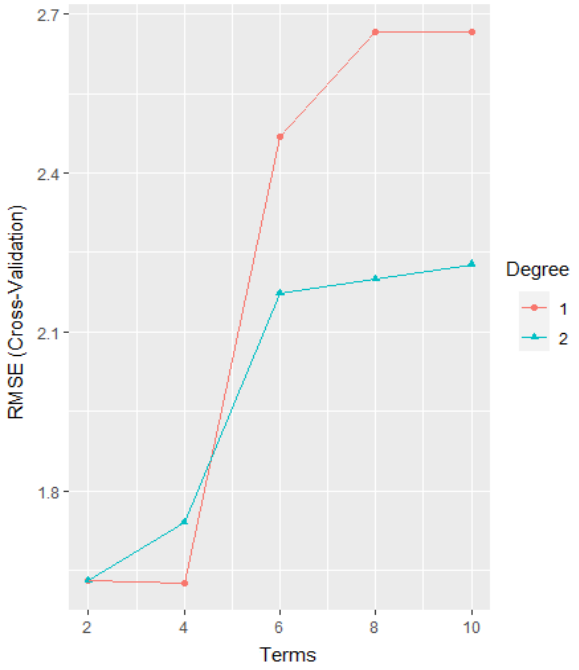

Figure 5: 10-fold CV RMSE for hyperparameter combinations in grid search.

Both the PCR and the MARS models also identified effective regression parameters. Table 1. summarizes the contributions of the indicator variables explored by the regression models together. According to Table 1, PM2.5 and NO_x have been recorded as the major predictor variables for CO_2 concentrations. Variable importance plots for each model structured in Figure 6. also indicates that there is no relationship between CO_2 emissions and non-methane volatile organic compounds.

Table 1: Contributions of indicator variables to models.

	% Importance	
Variable	**PCR**	**MARS**
PM2.5	100	100
NO_x	74.9	35.2
PM10	71.8	0.0
SOx	41.3	0.0
Volatile Organic Compounds	0.0	0.0

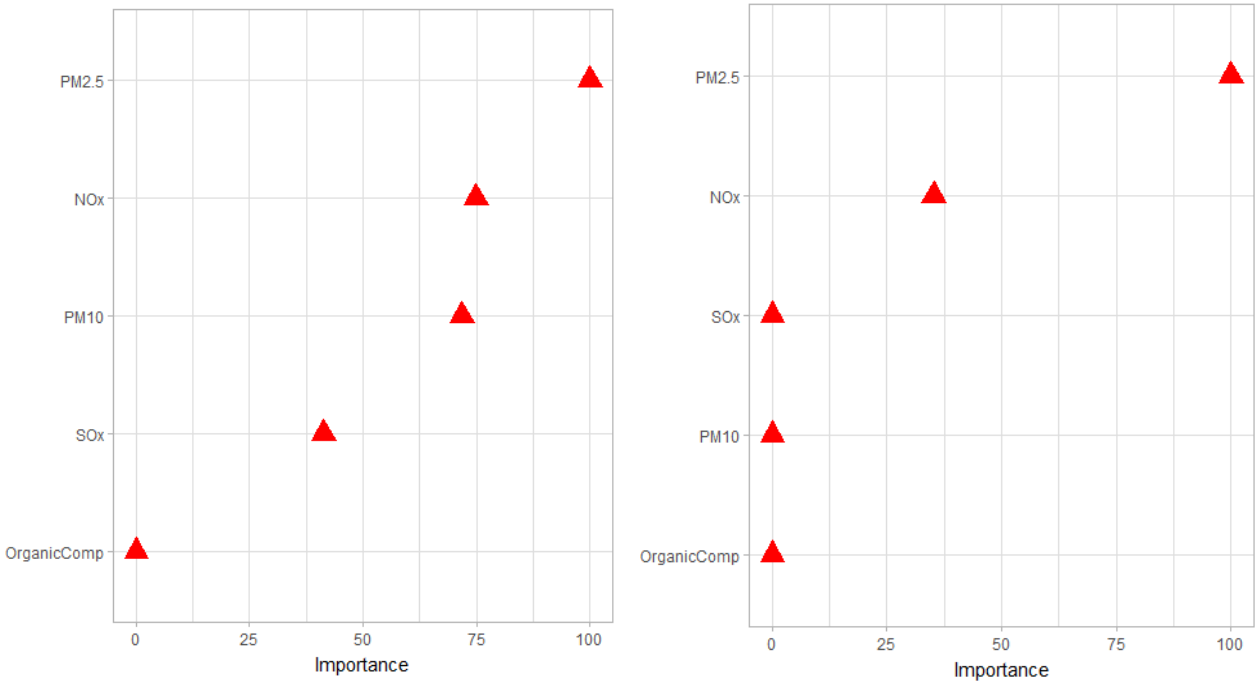

Figure 6: Variable importance for models - a) PCR, b) MARS

Along with the training implementations, the accuracy potentials of the regression models have also been inspected by a series of testing applications. To make a comparative evaluation, the conventional LSE-based multiple regression outcomes has also been added in the table. Mean Squared Error (MSE) has been considered as the performance measure index. The performances summarized in Table 2 revealed that testing performances of both the models outperform conventional regression. It should be underlined that the suggested models provided the high capabilities using limited number of variables. Therefore, these models have better generality, transparency as well as flexibility.

Table 2: Comparative performance analysis.

Regression Model	Training MSE	Testing MSE
MR	1.705	1.389
PCR	1.812	1.160
MARS	1.572	1.247

Looking at the outcomes at a close range, it should be noticed that there is a close relationship between fine particulate (tiny particles in the air that are of two and a half microns or less in width) matters (PM2.5) and carbon dioxide (CO2) concentrations. As a noticeable air pollutant, PM2.5 can be reaching the lungs. The main sources of these particles can be expressed as exhausts of the vehicles, burning of fuels and forest fires. CO_2 also comes from similar sources. Among the industries in Serbia, mining-quarrying and electricity-air conditioning supply have also critical importance on creation of emissions of both PM2.5 and CO_2 together. Therefore, a direct relationship can be recorded in the nature of things. According to measured data, a medium level correlation (about 60%) between these parameters was calculated by Pearson Coefficient Correlation. However, Table 1 remarks the contribution of PM2.5 to CO_2 estimations earth-shakingly (about 100%). Because both the models consider interactions and functional evaluations, the major importance of PM2.5 particles for the main greenhouse gas CO2 has been seen by these algorithms.

CONCLUSIONS

The focus has been on uncovering the relationships between CO_2 emitted into air from industries and air quality indicators in Serbia and they have been analyzed. To specify the dependencies between the major greenhouse gas (CO_2) and the major air quality indicators such as particulate matters (PM2.5, PM10), nitrogen and sulfur oxides (NO_x, SO_x), and volatile organic compounds, statistical learning-based evaluation tools were considered. To conduct realistic and multi-directional analyses, both parametric and non-parametric regression algorithms have been utilized.

The numerical experiments on training and testing data exhibited that both the modelling approaches have high and satisfactory accuracy capacities. The model interpretation studies revealed that the major pollutant to appraise CO_2 emissions is PM2.5. In addition, a medium level dependency between NO_x and CO2 has been recorded. Interpretable machine learning algorithms provide meta-data and new information for making black-box natural systems explainable. The findings provided a dependable statistical ground with valid algorithms and new information for the national authorities to appraise the relationships and sources and to make some precautions as a part of environmental management system of the country.

REFERENCES

Bal, F., & Vleugel, J. (2020). Towards more environmentally sustainable intercontinental freight transport. *International Journal of Transport Development and Integration*, *4*(2), 129–141. https://doi.org/ 10.2495/TDI-V4-N2-129-141

Bellinger, C., Jabbar, MSM., Zaiane, O., & Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*, 17, 907. https://doi.org/10.1186/s12889-017-4914-3

Boehmke, B., & Greenwell, B. (2020). *Hands-on machine learning with R*. Boca Raton: CRC Press.

Bollen, J., & Brink, C. (2014). Air pollution policy in Europe: Quantifying the interaction with greenhouse gases and climate change policies. *Energy Economics*, *46*(C), 202-215. https://doi.org/10.1016/j.eneco.2014.08.028

Brauer, M., Casadei, B., Harrington, R.A., Kovacs, R., & Sliwa, K. (2021). Taking a stand against air pollution – the impact on cardiovascular disease. *Global Heart*, *16*(1), 8. https://doi.org/10.1161/CIRCULATIONAHA.120.052666

Cristescu, T., Stoica, M.E., & Suditu, S. (2019). Research on the Carbon Dioxide Emission Factor as a Result of Fuel Combustion. *Revista De Chimie*, *70*(2), 585-590.

Gocheva-Ilieva, S.G., Ivanov, A.V., & Livieris, I.E. (2020). High performance machine learning models of large scale air pollution data in urban area. *Cybernetics and Information Technologies*, *20*(6), 49-60. https://doi.org/10.2478/cait-2020-0060

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning*: *Data mining, inference, and prediction*. Second Application, New York: Springer.

Liu, H-L., & Shen, Y-S. (2014). The impact of green space changes on air pollution and microclimates: A case study of the Taipei metropolitan area. *Sustainability*, *6*, 8827-55. https://doi.org/10.3390/su6128827

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: A review. *Front Public Health*, *8*, 14. https://doi.org/10.3389/fpubh.2020.00014

Martinez-Espana, R., Bueno-Crespo, A., Timn, I., Soto, J. Munoz, A., & Cecillia, J.M. (2018). Air-pollution prediction in smart cities through machine learning methods: a case of study in Murcia, Spain. *Journal of Universal Computer Science*, *24*(3), 261-276.

Melamed, M.L., Schmale, J., & von Schneidemesser, E. (2016). Sustainable policy-key considerations for air quality and climate change. *Current Opinion in Environmental Sustainability*, *23*, 85–91. https://doi.org/10.1016/j.cosust.2016.12.003

Ni, J.Q., Liu, S., Diehl, C.A., Lim, T.T., Bogan, B.W., Chen, L., Chai, L., Wang, K.Y., & Heber, A.J. (2017). Emission factors and characteristics of ammonia, hydrogen sulphide, carbon dioxide, and particulate matter at two high-rise layer hen houses. *Atmospheric Environment*, 154, 260-273. https://doi.org/10.1016/j.atmosenv.2017.01.050

OECD (2012). *OECD Environmental Outlook to 2050*. Organisation for Economic Co-operation and Development. https://download.repubblica.it/pdf/2016/ambiente/ocse-inazione.pdf

Özmen, A. (2016). *Robust optimization of spline models and complex regulatory networks.* Springer.

Paraschiv, S., & Paraschiv, L.S. (2020). Trends of carbon dioxide (CO_2) emissions from fossil fuels combustion (coal, gas and oil) in the EU member states from 1960 to 2018. *Energy Reports*, *6*(8), 237-242. https://doi.org/10.1016/j.egyr.2020.11.116

Statistical Office of the Republic of Serbia - Environment Database - SORS (2020). *Air Emissions accounts by NACE Rev.2 and for households*. https://data.stat.gov.rs/

Suryanarayana, T.M.V., & Mistry, P.B. (2016). *Principal component regression for crop yield estimation*. Singapore: Springer.

Tutmez, B. (2020). Air quality assessment by statistical learning-based regularization. *Cukurova University Journal of the Faculty of Engineering*, *35*(2), 271-278. https://doi.org/10.21605/cukurovaummfd.792412

Varmuza, K., & Filzmoser, P. (2009). *Introduction to multivariate statistical analysis in chemometrics*. CRC Press.

WHO (2019). *Health impact of ambient air pollution in Serbia: a call to action*. Copenhagen: The WHO Regional Office for Europe.

Zhang, W., & Goh, A.T.C. (2016). Multivariate adaptive regression splines and neural network models for prediction of pile drivability. *Geoscience Frontiers*, *7*, 45-52. https://doi.org/10.1016/j.gsf.2014.10.003

Zhang, X., Wang, X.Y., Bai, Z.P., & Han, B. (2013). Co-benefits of integrating PM10 and CO2 reduction in an electricity industry in Tianjin, China. *Aerosol and Air Quality Research*, *13*(2), 756-770. https://doi.org/10.4209/aaqr.2012.06.0144

TUMAČENJE ODNOSA ZAGAĐIVAČA I EMITOVANOG UGLJENDIOKSIDA U VAZDUH IZ INDUSTRIJA U SRBIJI

Problem zagađenja vazduha u Srbiji fokusira se na analizu odnosa između emitovane količine CO2 u vazduh iz industrija i indikatora kvaliteta vazduha kao što su čestice (PM2,5, PM10), oksidi azota i sumpora (NOk, SOk) i isparljiva organska jedinjenja. Da bi se identifikovale zavisnosti, uzeti su u obzir i parametarski i neparametarski statistički algoritmi za evaluaciju, zasnovani na učenju. Obe strukture modela dale su zadovoljavajuće procene s visokim nivoom tačnosti. Kao rezultat interpretacije modela, PM2,5 je zabeležen kao glavni indikator za istraživanje varijabilnosti koncentracija CO2. Implementacije su pokazale da mašinsko učenje koje se može tumačiti može da obezbedi metapodatke i dovoljno informacija da sistem kvaliteta vazduha crne kutije bude objašnjiviji. Samim tim, praktikovane alate za modeliranje, predstavljene međusobne relacije, kao i nove informacije, vlada može uzeti u obzir u okviru računarske strategije upravljanja životnom sredinom.

Ključne reči: Gas staklene bašte; Zagađenje vazduha; Statističko učenje; Regresija; Važnost promenljive.