Ozone Concentration Estimation from Infrared Images using Extinction Coefficient

Alexandra Duminil[1][0000−0001−7180−4119], Jean-Philippe Tarel[1][0000−0002−9241−5347], and Jean Dumoulin[2][0000−0003−0378−6136]

[1] University Gustave Eiffel, COSYS-PICS-L, F-77454 Marne-la-Vallée, France
[2] University Gustave Eiffel, Inria, COSYS-SII, I4S Team, F-44344 Bouguenais, France jean-philippe.tarel@univ-eiffel.fr

Abstract. Air quality assessment requires concentration measurement of various polluting gases, typically requiring multiple expensive sensors, each dedicated to a specific pollutant. Therefore, there is a need for cost-effective sensors capable of detecting one or more pollutants, with less reliability, such as camera-based sensors, but enabling denser sampling. In this paper, we investigate how the extinction coefficients estimated from an infrared camera may be useful for predicting ground-level ozone concentration. In addition to these coefficients, we show how weather and pollution measures collected from stations near the studied area are useful, with different machine learning methods, to better predict ozone concentration. The performance of the models is validated through a comprehensive evaluation using MAE, RMSE and R-Squared metrics. Parameter selection methods are also used to study the impact of different meteorological parameters and other pollutant concentrations on the prediction of ozone concentration.

Keywords: Pollution · Ozone prediction · Infrared images · Visibility

1 Introduction

Ambient ozone (O_3) is a naturally occurring gas in the atmosphere, formed through the chemical reaction of primary pollutants (nitrogen dioxide (NO2), volatile organic compounds (VOCs)) under ultraviolet radiation. However, it poses toxic risks to living organisms. Under persistent anticyclonic conditions, characterized by dry, sunny weather and low wind activity, it tends to gradually accumulate in the atmosphere [1]. As a result, implementing O_3 monitoring and prevention systems could significantly benefit public health. Monitoring air quality typically relies on measuring the concentration of various polluting gases, which is usually done using multiple expensive sensors, each dedicated to a specific pollutant. Therefore, there is a growing need for low-cost effective solutions, such as camera-based sensors, capable of detecting one or more pollutants, enabling more extensive and denser sampling. Several models have been proposed in the literature to estimate the concentration of pollutants and IQAs in particular, including statistical, deterministic, physically-based, and Machine

Learning (ML)-based models. According to Kumar et al. [9], ML models have been shown to be more efficient than statistical and deterministic models. In particular, SVR [18], Random Forest [2], K-means, multiple linear regression, and MLP models are widely used. Liang et al. [10] investigated the performance of six ML classifiers in predicting Taiwan's IQA based on 11 years of data. The authors reported, in this study, that Adaptive Boosting (AdaBoost) and Stacking Ensemble were the most suitable, but prediction performance seems to vary between geographical regions. Several studies [12, 17] have shown that the performance of prediction models is improved by adding meteorological data such as temperature, humidity, wind speed and direction. In [8], the linear regression model, which is a simple model, is not effective in taking into account local variations. Juarez et al. [7] proposed a comparison analysis of O_3 prediction in Delhi, India, using eight different ML methods. The study, carried out both seasonally and annually, showed improvements in $O3$ prediction, particularly during winter. Deep CNNs are not widely used for pollutant prediction but have been shown to be suitable for non-linear processes related to air quality [4].

Some studies proposed image processing-based methods, which focused on particulate matter concentration predictions [11, 16, 3, 14]. Part of them aimed to analyze images by focusing on sky regions and estimating visual features such as the dark channel prior [11, 16], or the color histogram [3]. Mohan et al. [14] proposed using pre-trained CNN methods to directly classify outdoor RGB images with different levels of pollution.

While previous studies have proposed image-based methods to predict particulate matter concentrations from RGB images, our work aims to go a step further by using infrared images to help estimate ozone concentrations in Champs-sur-Marne, a town around twenty kilometers east from Paris, France. Specifically, our work stands out by investigating the interest in using extinction coefficients k calculated from infrared camera images, a specific feature, jointly with other meteorological and pollution measures. To the best of our knowledge, this approach has not yet been investigated in the literature. We also studied the interest of different input measures with also their lagged values and the interest of using a parameter selection method such as SHAP. Different ML models are used to perform predictions including SVR, random forest, boosting, MLP and CNN.

2 Materials and methods

2.1 Data

The data are collected from our system located on the roof of the University, at Champs-sur-Marne in France. It is composed of two RGB cameras, an infrared camera and a visibilimeter, which allow us to obtain meteorological data such as temperature, background luminance, cumulative water sum (cws) and the visibility distance. Pollution data, which are used to obtain the ground truth, are collected from the Airparif database [1] until 2018. Currently, the dataset comprises a single scene, as obtaining data over a long period of time using a specific sensor system is challenging. However, even though the dataset only

contains a single scene, the system takes pictures continuously, enabling data to be obtained under different meteorological conditions.

The dataset used in this study consists of a data series of 18500 samples collected between 2018 and 2023. Data for 2020 were excluded due to a lower sample count compared to other years. To avoid imbalances, the number of samples for each remaining year is approximately equal. Then, missing and NaN values led to the removal of the data line.

2.2 Method

Air quality depends primarily on the intensity of pollutant emissions, which are influenced by various factors, particularly meteorological conditions. These conditions also determine the dispersion or accumulation of pollutants in the atmosphere. Wind and rain, for instance, promote the dispersion, mixing, and removal of pollutants. In contrast, persistent anticyclonic conditions, characterized by dry, sunny weather and the absence of wind at ground level, can lead to gradual accumulation of pollutants. Consequently, the summer season is characterized by hot and sunny days that favor the formation of ozone. Fig. 1, shows the O_3 concentration over several years. A low concentration of O_3 is noticeable in autumn and winter, while it increases in spring and summer.

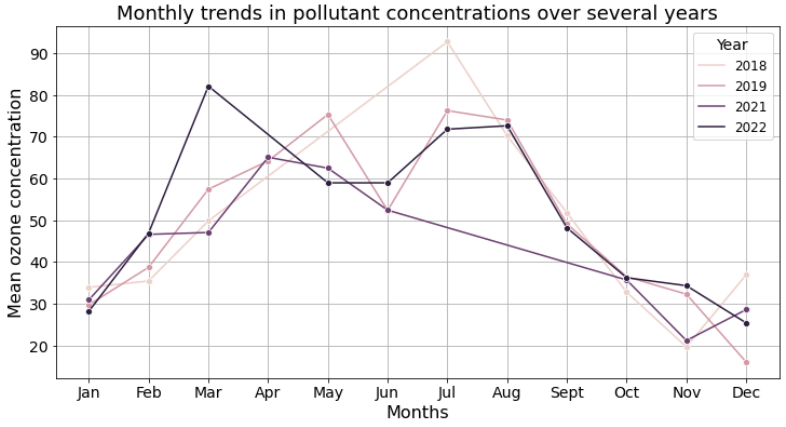

Fig. 1. Lines plot describing the monthly trends in O_3 concentrations over several years.

In addition to the meteorological data, we aim to exploit the infrared images provided by the camera. O_3 absorbs and emits in the infrared range and is strongly influenced by meteorological conditions, including sunshine and high temperatures. Fig. 2 presents cropped images from the infrared camera with associated O_3 concentration. The right image shows higher contrast, likely due to seasonal temperature differences. The dark area, where the horizon is barely

visible, may result from atmospheric absorption, where ozone would be involved. Infrared imagery can be used to visualize thermal variations in the atmosphere, leading indirectly to the location of ozone.

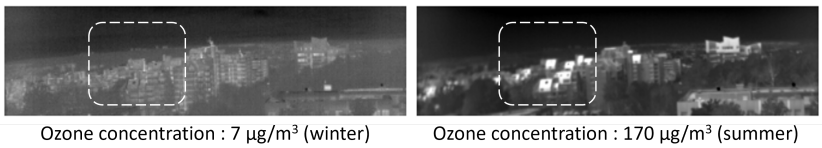

Fig. 2. Infrared camera images of areas of interest with associated O_3 concentration.

Extinction coefficients *k* To achieve this, we rely on the use of extinction coefficients, which have been used in previous works to estimate the concentration of particulate matter in the atmosphere [11]. The visible veil, caused by high concentrations of particulate matter, is due to the interaction of light with the particles in the air, called light scattering, which causes an attenuation of transmitted light. The Koschmieder law [13] is a physically based model of the apparent luminance of objects on sky background near the horizon:

$$I = I_0 e^{-kd} + I_s(1 - e^{-kd}) \tag{1}$$

where I is the intensity of an object seen at distance d, I_0 is the clear intensity of this object, I_s is the intensity of the sky and k is the atmospheric extinction coefficient to estimate. Eq. 1 assumes constant lighting conditions. In practice, the weather and the position of the Sun vary between the days and seasons. For this reason, assuming a cold sky, the hypothesis of neglecting the second term of I in Eq. 1 is explored, leading to a simpler model with only the first term:

$$I \approx I_0 e^{-kd} \tag{2}$$

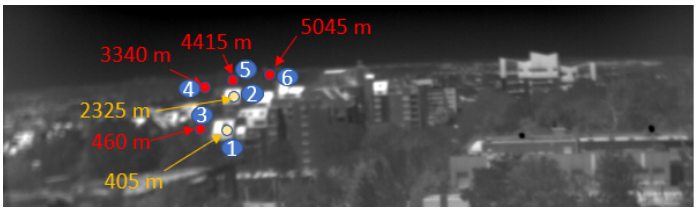

Fig. 3. Infrared image with distances in meters and points of interest with marks used for coefficient calculations.

Estimated k	Target object	Set of point marks
k_{optim}	vegetation	3,4,5
$k1_{\text{b}}$	buildings	1,2
$k2_{\text{v}}$	vegetation	3,4
$k3_{\text{v}}$	vegetation	3,5
$k4_{\text{v}}$	vegetation	3,6
$k5_{\text{v}}$	vegetation	4,6

Table 1. Label names of the estimated extinction coefficient k depending on the set of used point marks, with numbers as shown in Fig. 3.

Fig. 3 presents the different points of interest with a number mark and displays their approximate distances to the camera in meters (m). Table 1 shows the label names of the estimated extinction coefficient k depending on the used point marks shown in Fig. 3. Points of interest were chosen on buildings (1 and 2) and on vegetation (3 to 6). The purpose is to investigate the optimum choice of the point marks set. Buildings tend to get warmer compared to vegetation. Thus, it could be relevant to analyze the impact of these choices on the calculation of the extinction coefficient. When considering the simpler model of Eq. 2, the k can be computed using only two image points at different known distances to simplify the unknown I_0 [5]. Therefore, k_n can be estimated as follows:

$$k_n = \frac{1}{(d_j - d_i)} log(\frac{I_i}{I_j}) \tag{3}$$

where I_i and I_j are the average intensities of a 5×5 area around points i and j, d_i and d_j are the distances between camera and pointed objects i and j (see Fig. 3). The k_n computations are independent of the global illumination conditions.

When considering the complete model of Eq. 1, the equation is with three unknowns: k, I_s and I_0, and thus three points at three different distances are requested to estimate k. These unknowns' values are obtained by least squares minimization. The k_{optim} can thus be computed from where the minimum of the following error is achieved:

$$\begin{aligned} &(I_3 - (I_{03}\, e^{-kd_3} + I_s(1 - e^{-kd_3}))^2 + \\ &(I_4 - (I_{03}\, r_{34}\, e^{-kd_4} + I_s(1 - e^{-kd_4}))^2 + \\ &(I_5 - (I_{03}\, r_{35}\, e^{-kd_5} + I_s(1 - e^{-kd_5}))^2 \end{aligned} \tag{4}$$

where r_{34} and r_{35} are the ratios of the mean intensities between points 3 and 4 and points 3 and 5, respectively. By comparing these different estimates, this will allow us to determine what is the best estimate of the extinction coefficient of O_3 concentration.

Data analysis Studies have shown that pollutant concentrations over time t depend on previous values [6]. To analyze these temporal dependencies, the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)

are used. ACF captures the general correlation with past values, while PACF isolates the direct effect of each lag. The time increment in our case is one hour. As with pollutants, meteorological data and the k coefficients at a given time t are correlated with the two previous hours. These additional information are also used as input parameters. The suffixes $lag1$ and $lag2$ are used to differentiate time-shifted parameters in the rest of the paper.

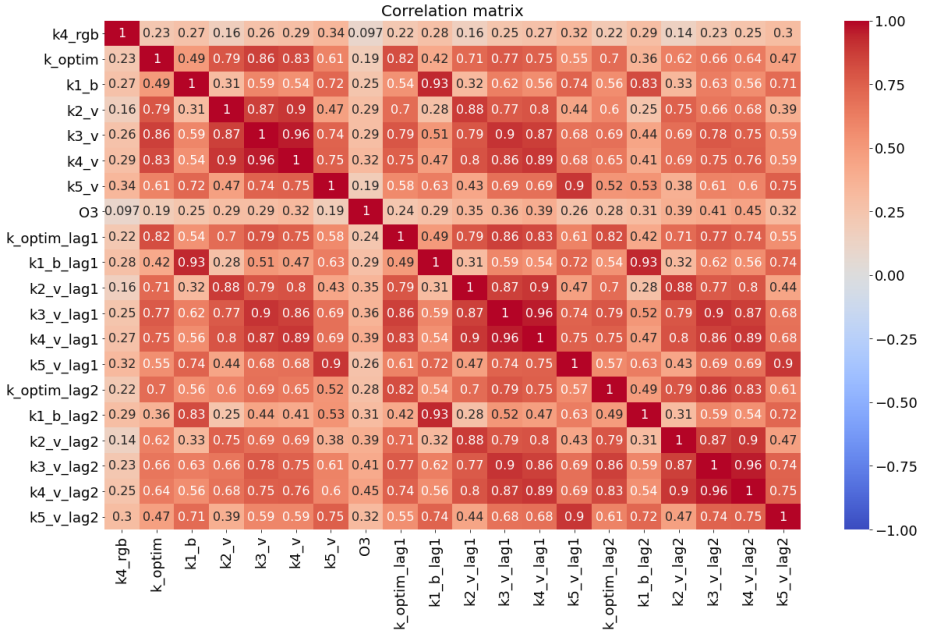

Fig. 4. Heatmap of the linear correlation matrix between O_3 and the different extinction coefficients.

Fig. 4 presents the heatmap of the correlation matrix obtained from a set of data including the estimated k coefficients, and the Airparif data O_3 . The values turn red as positive correlations increase and blue for negative correlations. The weakest correlations with O_3 are k_{optim} and $k5_v$ followed by $k1_b$. A similar pattern was observed in the lagged versions, which tended to increase with lag. This suggests that these k values may not be optimal for estimating ozone concentration. For $k1_b$, the points of interest taken in the buildings may slightly distort the results due to their higher temperature. For $k5_v$, the distance between the two points is around 1705 m, corresponding to the shortest case. The best correlation values with O_3 are $k4_v$ and its lagged values. The points of interest were chosen in vegetation areas and at a greater distance than the other k values. To compare with the extinction coefficient obtained in the visible spectrum, the parameter $k4_{rgb}$ has been estimated similarly to $k4_v$ but on RGB

images and is added at the first line of the correlation matrix. Its correlation value with O_3 is lower than 0.1, showing the interest of using infrared images instead of RGB. However, the correlation matrix captures only linear correlations. Thus, in addition, the mutual information (MI) between O_3 and the different extinction coefficients is computed to measure a non-linear link between them. The MI is presented in Fig. 5. The results confirm previous statements about the correlations of k_{optim}, $k5_v$ and $k1_b$ and their lagged values. In general, the coefficients with the highest MI scores are $k3_v$ and $k4_v$ with their lags.

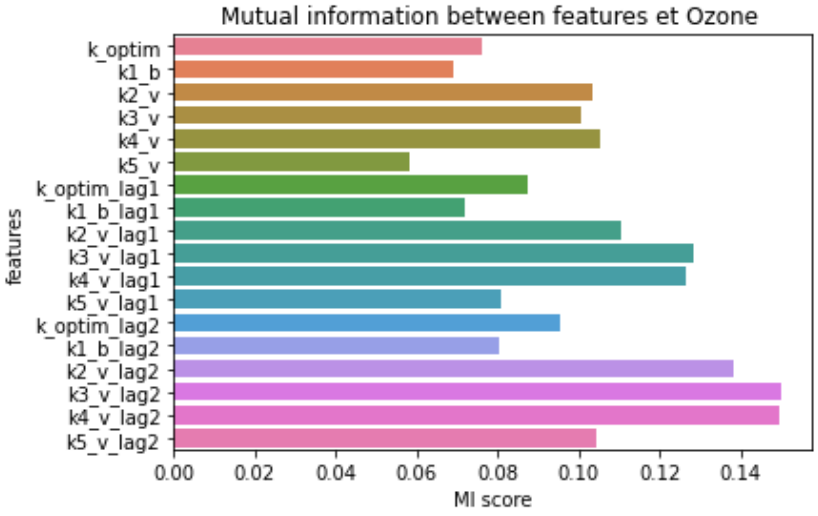

Fig. 5. Bar plot describing the mutual information between O_3 and the different extinction coefficients.

Meteorological data The use of meteorological data has been shown to be essential to predict or monitor the concentration of pollutants. Many studies have used data such as absolute humidity, air temperature, absolute pressure, and wind direction. In our study, our objective is to use the available meteorological data. The meteorological sensor is located on the station with the cameras on our acquisition site. The data retrieved are air temperature, background luminance, cumulative water sum (cws) and visibility distance. The background luminance corresponds to the ambient light and the cws is the total accumulated precipitation. The visibility distance may depend on particles in the air, rain or fog. Ozone concentration peaks appear in cycles, as shown in Fig. 1. The peaks occur at a certain time of year, mainly in summer and generally in the afternoon. It thus seems relevant to add hours value, days value and seasons value as cyclical variables to the previous meteorological data vector. For example, the variable

day is included as two variables: $sin_{day} = \frac{sin(2\pi \times day)}{7}$ and $cos_{day} = \frac{cos(2\pi \times day)}{7}$. The number 7 denotes the number of days in a week. This is performed similarly for hours and seasons.

The concentration of pollutant from Air-Quality stations close to Champs-Sur-Marne are also interesting to study. Fig. 6 presents the MI with the different parameters previously selected and the O_3 concentration. We observed a strong link between $NO2$ and O_3. This can be due to the formation of O_3 by chemical reaction with primary pollutants, including $NO2$. Temperature and cws are also important features. The former is indeed essential for the formation of O_3, since its concentration increases with temperature. For the latter, precipitation favors the dispersion and leaching of pollutants. As a result, the O_3 concentration should decrease as cws increases. The importance of particulate matter PM2.5 and PM10 can be noticed. In fact, there are involved in slowing the fall of radical aerosols, which are one of the precursors of O_3 [15]. Some parameters, such as luminance, have more information in the lag2 version. Cyclic parameters appear less important, and MI between O_3 and the other data varies across features. Since some parameters may be redundant, it is essential to select the most predictive ones.

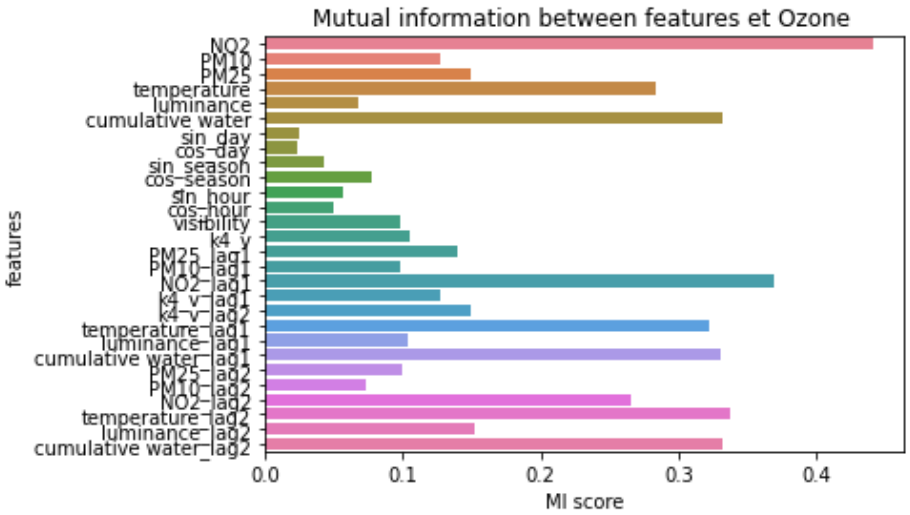

Fig. 6. Bar plot describing the mutual information between O_3 concentration and other selected parameters.

3 Results

The collected data serves to train different ML models in order to predict O_3 concentration. Inspired by the state-of-the-art, classical ML models are selected:

Linear regression (LR), SVR, Random forest (RF), Xtreme gradient boosting (XB), as well as MLP and CNN. The CNN model consists of five layers of 1D convolution followed by two dense layers. The original data consists of a data series of approximately 18500 collected data from 2018 to 2023. The ML models are trained on data of size 11863 and 2966 for validation, including data from 2018 to 2023. The test sets $t1$ and $t2$ are based on data from 2022 and 2023, respectively. The data were split using odd months for training/validation and even months for testing, ensuring test independence while retaining data. Three performance metrics are used in this section: The Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-Squared (R^2), to measure the prediction errors for O_3 concentration prediction. Lower MAE and RMSE values and higher R^2 value demonstrate better performance.

	Without k parameters + lags						**With k parameters + lags**					
Methods	LR	SVR	RF	XB	MLP	CNN	LR	SVR	RF	XB	MLP	CNN
MAE v	0.436	0.250	0.276	0.280	0.278	**0.188**	**0.432**	**0.234**	**0.270**	**0.280**	**0.270**	0.190
RMSE v	0.554	0.336	0.364	0.381	0.367	**0.255**	**0.548**	**0.315**	**0.358**	**0.377**	**0.350**	0.257
R^2 v	0.707	0.892	0.873	0.862	0.871	**0.938**	**0.714**	**0.905**	**0.878**	**0.864**	**0.883**	0.937
MAE $t1$	0.503	0.450	0.396	0.436	**0.412**	0.423	**0.490**	**0.448**	**0.389**	**0.427**	0.433	**0.399**
RMSE $t1$	0.615	**0.577**	0.504	0.554	**0.529**	0.544	**0.599**	0.581	**0.496**	**0.545**	0.557	**0.510**
R^2 $t1$	0.637	**0.681**	0.756	0.706	**0.732**	0.716	**0.656**	0.676	**0.764**	**0.715**	0.702	**0.751**
MAE $t2$	**0.473**	**0.398**	0.349	0.376	**0.384**	0.395	0.479	0.409	**0.342**	**0.365**	0.400	**0.375**
RMSE $t2$	**0.590**	**0.514**	0.453	0.485	**0.498**	0.502	0.598	0.534	**0.448**	**0.472**	0.522	**0.492**
R^2 $t2$	**0.696**	**0.769**	0.821	0.795	**0.783**	0.780	0.688	0.751	**0.825**	**0.805**	0.762	**0.788**

Table 2. Prediction errors on validation (v) and test data ($t1$ and $t2$). The best values are in bold.

Table 2 presents the prediction errors for validation and tests with and without k parameters for different ML models. The general results on validation v and on tests $t1$ and $t2$ demonstrate better performance with the k parameters, compared to without. Specifically, the CNN, RF and Boosting models produce the best results on $t2$ and $t3$, which include the k parameters. Above all, RF appears to be the most suitable model for this type of task.

Based on these results, a SHAP analysis is applied on the CNN model to display the importance of each parameter, using the SHAP summary plot in Fig. 7. As we saw in the previous section, NO_2 has a significant impact on O_3 concentration. More specifically, in terms of SHAP values, higher NO_2 values tend to reduce O_3 levels. This means that high NO_2 levels are not required to raise O_3, as they also correlates with meteorological factors such as temperature, especially at night and during the colder seasons. In addition, the plot shows that higher temperatures, luminance and k_4 values are linked to increased ozone concentrations, as indicated by their positive SHAP values. The visibility feature has negative values associated with a decrease in O_3 concentration. Finally, the others' values demonstrate their limited impact on O_3 concentration. In sum-

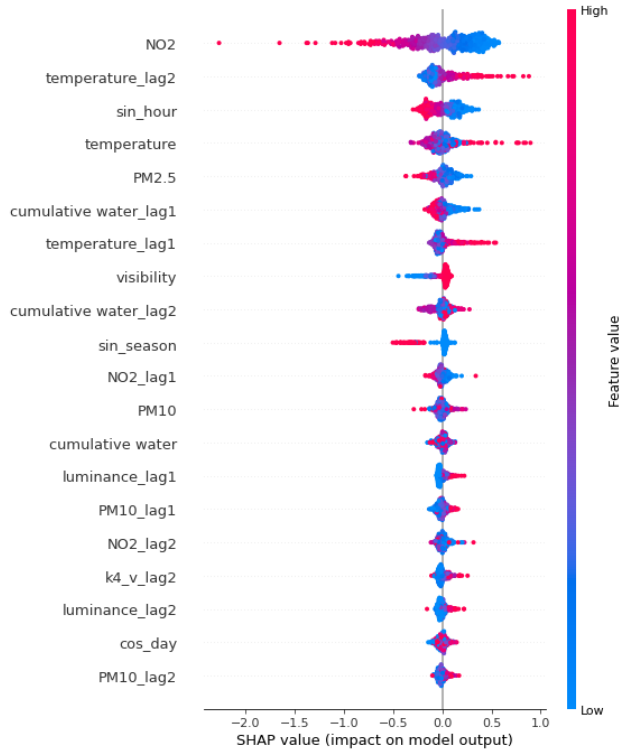

Fig. 7. Graph showing the importance of parameters using the SHAP method.

mary, the use of the SHAP method provides an overview of the impact of the various parameters on the learned model. However, the SHAP method appears to be relatively sensitive, since the results differ significantly depending on the types of model, the size of the datasets, and the number of input parameters.

4 Conclusion

The purpose of this work was to study the interest in using extinction coefficients computed from infrared images independently of their acquisition time to help predict O_3 concentration. The experimental study is conducted in Champs-sur-Marne, France, from a dataset collected quite regularly each hour from 2018 to 2023 under various different weather and traffic conditions. The first experiments are based on data from a single scene acquired under different weather conditions and throughout the day over several years. The preliminary results showed a significant correlation between these extinction coefficients and the O_3 concentration, making the use of camera data interesting for O_3 prediction. We were also able to measure the impact of other meteorological conditions that are relatively rarely studied in the literature, such as visibility or sky luminance.

Then, we performed a comparison with several ML models and a CNN, which gave promising results on test sets. The perspective of collecting more data over the years to cover as many situations as possible and to improve the generalization properties of prediction models is important to consider. In future work, we also aim to extend the proposed method to more infrared scenes, improve the quantitative evaluation by adding advanced models, and to exploit the use of multi-spectral cameras.

Acknowledgment

Thanks to BRIGHTER project for funding. BRIGHTER has received funding from the Chips Joint Undertaking (JU) under grant agreement No 101096985. The JU receives support from the European Union's Horizon Europe research and innovation program and from France, Belgium, Portugal, Spain and Turkey.

References

1. Airparif: L'observatoire de la qualité de l'air en île-de-france (2023), https://www.airparif.fr/, consulted on 16 April 2025
2. Chen, G., Chen, J., Dong, G.h., Yang, B.y., Liu, Y., Lu, T., Yu, P., Guo, Y., Li, S.: Improving satellite-based estimation of surface ozone across China during 2008–2019 using iterative random forest model and high-resolution grid meteorological data. Sustainable Cities and Society **69**, 102807 (2021)
3. Enggari, S., Tajuddin, M., et al.: Identification of air pollution levels based on sky image using color histogram method. In: 2023 Sixth International Conference on Vocational Education and Electrical Engineering (ICVEE). pp. 77–83. IEEE (2023)
4. Eslami, E., Choi, Y., Lops, Y., Sayeed, A.: A real-time hourly ozone prediction system using deep convolutional neural network. Neural Computing and Applications **32**, 8783–8797 (2020)
5. Hautiere, N., Aubert, D., Dumont, , Tarel, J.P.: Experimental validation of dedicated methods to in-vehicle estimation of atmospheric visibility distance. IEEE Transactions on Instrumentation and Measurement **57**(10), 2218–2225 (2008). https://doi.org/10.1109/TIM.2008.922096
6. Jairi, I., Ben-Othman, S., Canivet, L., Zgaya-Biau, H.: Enhancing air pollution prediction: A neural transfer learning approach across different air pollutants. Environmental Technology & Innovation **36**, 103793 (2024)
7. Juarez, E.K., Petersen, M.R.: A comparison of machine learning methods to forecast tropospheric ozone levels in Delhi. Atmosphere **13**(1), 46 (2021)
8. Jumin, E., Zaini, N., Ahmed, A., Abdullah, S., Ismail, M., Sherif, M., Sefelnasr, A., el Shafie, A.: Machine learning versus linear regression modelling approach for accurate ozone concentrations prediction. eng appl comput fluid mech 14: 713–725 (2020)
9. Kumar, K., Pande, B.: Air pollution prediction with machine learning: a case study of indian cities. International Journal of Environmental Science and Technology **20**(5), 5333–5348 (2023)
10. Liang, Y.C., Maimury, Y., Chen, A.H.L., Juarez, J.R.C.: Machine learning-based prediction of air quality. applied sciences **10**(24), 9151 (2020)

11. Liu, C., Tsow, F., Zou, Y., Tao, N.: Particle pollution estimation based on image analysis. PloS one **11**(2), e0145955 (2016)
12. Madan, T., Sagar, S., Virmani, D.: Air quality prediction using machine learning algorithms–a review. In: 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN). pp. 140–145. IEEE (2020)
13. Middleton, W.E.K.: Vision through the atmosphere. In: geophysik ii/geophysics ii, pp. 254–287. Springer (1957)
14. Mohan, A.S., Abraham, L.: Hybrid transfer learning approach using multiple pretrained models for classification of outdoor images into AQI classes. In: 2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE). vol. 7, pp. 283–288. IEEE (2022)
15. Qu, Y., Wang, T., Cai, Y., Wang, S., Chen, P., Li, S., Li, M., Yuan, C., Wang, J., Xu, S.: Influence of atmospheric particulate matter on ozone in Nanjing, China: Observational study and mechanistic analysis. Advances in Atmospheric Sciences **35**, 1381–1395 (2018)
16. Samsami, M., Shojaee, N., Savar, S., Yazdi, M.: Classification of the air quality level based on analysis of the sky images. In: 2019 27th Iranian Conference on Electrical Engineering (ICEE). pp. 1492–1497. IEEE (2019)
17. Simu, S., Turkar, V., Martires, R., Asolkar, V., Monteiro, S., Fernandes, V., Salgaoncary, V.: Air pollution prediction using machine learning. In: 2020 IEEE Bombay Section Signature Conference (IBSSC). pp. 231–236. IEEE (2020)
18. Su, X., An, J., Zhang, Y., Zhu, P., Zhu, B.: Prediction of ozone hourly concentrations by support vector machine and kernel extreme learning machine using wavelet transformation and partial least squares methods. Atmospheric Pollution Research **11**(6), 51–60 (2020)