Air Quality, Atmosphere & Health (2019) 12:1049–1057
https://doi.org/10.1007/s11869-019-00721-9

Macao air quality forecast using statistical methods

Man Tat Lei[1,2] · Joana Monjardino[3] · Luisa Mendes[1] · David Gonçalves[2] · Francisco Ferreira[3]

Received: 21 May 2019 / Accepted: 11 July 2019 / Published online: 27 July 2019

Abstract

The levels of air pollution in Macao often exceeded the levels recommended by WHO. In order for the population to take precautionary measures and avoid further health risks under high pollutant exposure, it is important to develop a reliable air quality forecast. Statistical models based on linear multiple regression (MR) and classification and regression trees (CART) analysis were developed successfully, for Macao, to predict the next day concentrations of NO_2, PM_{10}, $PM_{2.5}$, and O_3. All the developed models were statistically significantly valid with a 95% confidence level with high coefficients of determination (from 0.78 to 0.93) for all pollutants. The models utilized meteorological and air quality variables based on 5 years of historical data, from 2013 to 2017. Data from 2013 to 2016 were used to develop the statistical models and data from 2017 was used for validation purposes. A wide range of meteorological and air quality variables was identified, and only some were selected as significant independent variables. Meteorological variables were selected from an extensive list of variables, including geopotential height, relative humidity, atmospheric stability, and air temperature at different vertical levels. Air quality variables translate the resilience of the recent past concentrations of each pollutant and usually are maximum and/or the average of latest 24-h levels. The models were applied in forecasting the next day average daily concentrations for NO_2 and PM and maximum hourly O_3 levels for five air quality monitoring stations. The results are expected to be an operational air quality forecast for Macao.

Keywords Particulate matter · PM2.5 · PM10 · NO2 · O3 · Macao

✉ Man Tat Lei
l.tat@campus.fct.unl.pt; lei.man.tat@usj.edu.mo

Joana Monjardino
jvm@fct.unl.pt

Luisa Mendes
lc.mendes@fct.unl.pt

David Gonçalves
david.goncalves@usj.edu.mo

Francisco Ferreira
ff@fct.unl.pt

[1] Department of Sciences and Environmental Engineering, NOVA School of Science and Technology, NOVA University Lisbon, Lisbon, Portugal

[2] Institute of Science and Environment, University of Saint Joseph, Macau, China

[3] Center for Environmental and Sustainability Research, NOVA School of Science and Technology, NOVA University Lisbon, Lisbon, Portugal

Introduction

Seven million people die every year from the effects of air pollution. More than 90% of such deaths are in developing countries (WHO 2019). Across southern Asia, levels of fine particulate matter ($PM_{2.5}$) and surface ozone (O_3) exceed the World Health Organization (WHO) limits for much of the year (Kumar et al. 2018). Macao is located in Southern China, in the Pearl River Delta (PRD) region. The levels of nitrogen dioxide (NO_2), particulate matter (PM), particulate matter with an average aerodynamic diameter below 10 μm and 2.5 μm (PM_{10} and $PM_{2.5}$, respectively), and ozone (O_3) in Macao are high and often exceed the established limit values recommended by WHO's air quality guidelines (AQG). Since 2010, the worst air quality index classes in Macao have been due to PM_{10} and $PM_{2.5}$ (SMG 2019). Macao was listed as the number one most densely populated region in the world (Sheng and Tang 2013), with a population density of about 20,000 inhabitants/km^2. A significant proportion of Macao urban population is being exposed to air pollutant concentrations above the limit or target values.

The exposure to air pollutants such as NO_2, PM, and O_3 increase the chance of hospital admissions for cardiovascular and respiratory disease and mortality in the world (Liu and Peng 2018; WHO 2018). O_3 at the ground level is associated with numerous harmful effects on respiratory health, at levels commonly found in urban areas throughout the world, contributing to morbidity and hospital admissions related to respiratory disease, even at low ambient levels (Entwistle et al. 2019). Regarding particulate matter, for human health, small particles ($PM_{2.5}$) are particularly dangerous as they can penetrate deeply into the lungs and be transported directly into the bloodstream (Wiśniewska et al. 2019). Furthermore, mixtures of NO_2-$PM_{2.5}$-O_3 exist in ambient environments, being the combinations of these pollutants more harmful to human health (a mixture with relatively low levels of some pollutants combined with relatively high levels of other pollutants was found to be equally or more harmful than a mixture with high levels of all pollutants) (Liu and Peng 2018). In Macao, traffic-related pollution is high, primarily due to high vehicle emissions and urban canyon topology (He et al. 2000).

In this context, it is relevant to develop a reliable methodology to forecast the concentration of air pollutants, which can provide an alert for health hazards in advance, in a way that the population can take precautionary actions to avoid exposure.

Recent studies have been conducted to access meteorological influence on air quality (Tong et al. 2018a, b; Xie et al. 2019), and related to air quality forecast (Lee et al. 2017; Deng et al. 2018), both in PRD region. The current paper focuses the development of air quality forecast models by statistical methods for the most critical air pollutants in Macao.

The methods for the prediction of the air pollutant concentration can be roughly divided into two types: deterministic and stochastic. Statistical approach learns from historical data and predicts the future behavior of the air pollutants. Meteorological conditions significantly affect the levels of air pollution in the urban atmosphere, due to their important role in the transport and dilution of pollutants. It has also been concluded that there is a close relationship between the concentration of air pollutants and meteorological variables (Zhang and Ding 2017). Thus, multiple linear regression models (MR) are trained based on existing measurements and are used to predict concentrations of air pollutants in the future, according to the corresponding meteorological variables.

The Greater Bay Area (GBA) of China consists of nine cities of Guangdong province, and the Special Administrative Region of Hong Kong and Macao. The synoptic situation of Macao and other cities of the GBA is closely related due to its geographic proximity. The GBA experiences a complex temporal and spatial climatic condition due to topographic variations, urban morphology, and land-water contrasts. Located along the southeast coast of Mainland China, Macao is surrounded by the sea on three sides, with a subtropical oceanic monsoon climate that is characterized by high temperatures, high rates of evaporation, high levels of atmospheric moisture, and abundant rainfall (SMG 2014). In winter, Macao is influenced by the north monsoon, the climate is cold and dry with the predominant wind from the north quadrant. In summer, the northeast monsoon is replaced by the strong southwest monsoon with heavy rains. Spring and autumn are transition periods.

Recent studies (Tong et al. 2018a, b) showed a rise of surface temperature and a drop of surface absolute humidity and wind speed at GBA due to the decline of vegetation and irrigated cropland. The landscape of GBA is characterized by a large flatland surrounded by the Nanling Mountains which can prevent air pollution from the central part of China reaching the GBA. Nevertheless, the northeast monsoon present during the winter may transport pollutants from northern and eastern China, along the coastline to the region of GBA (Tong et al. 2018a, b). PM levels are usually measured higher during the winter season, from December to February, due to the northern wind, bringing the air pollutants to the region, lowering mixing height, and fewer amount and lower frequency of rainfall. During summer season, from June to August, PM levels are usually measured lower due to the southern winds from the China sea, higher mixing height, higher frequency, and amount of rainfall, which allow for a better air pollution dispersion and deposition conditions (Lopes et al. 2016).

The air pollution of the GBA is normally associated with emission sources at alternating spatial scales from local to regional and transboundary (Tong et al. 2018a), under certain synoptic conditions. Estimates show that, in this region, for nitrogen oxides (NO_x), mobile sources account for the majority of emissions (50%). For PM, the industrial sector is the main emitter, followed by mobile sources (Zheng et al. 2009). O_3 is not emitted directly to the atmosphere, but is formed in reactions between NO_x and volatile organic compounds (VOC), being these reactions driven by absorbed solar radiation (Reid et al. 2008).

Materials and methods

The statistical methods selected for this paper were both multiple linear regression analysis (MR) and classification and regression tree (CART). Those can be a useful and straightforward tool in air quality studies (Choi et al. 2013; Martinez et al. 2018; Cassmassi 1987; Clapp and Jenkin 2001). As one of the advantages of the CART analysis is its effectiveness in explaining the variations in pollutant levels solely by a combination of meteorological conditions, regression trees can identify specific meteorological conditions that lead to low or elevated pollutant concentrations (Choi et al. 2013). The basic concept of the CART approach is to make a hierarchy of

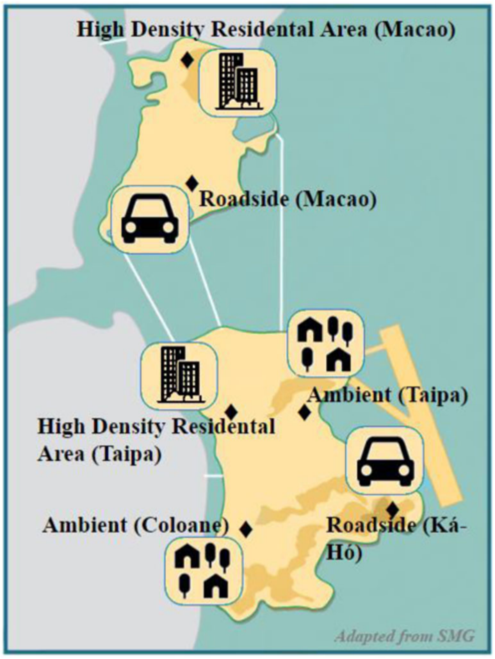

Fig. 1 Air quality monitoring network spatial location in Macao

binary decisions, each of which splits distribution/variation of a target variables into two mutually exclusive branches (groups) based on the explanatory variable/value showing the largest reduction in variations in target variable after the split (Choi et al. 2013).

Following precedent experiences (Cassmassi 1987; US EPA 2003; Durão et al. 2016; Oduro et al. 2016), the statistical models were initially created using MR analysis. As an approach to obtain improved results, mainly regarding a better prediction of high pollutant levels, the CART analysis was chosen to better predict the maximum concentrations.

Statistical models, based on MR and CART, were applied to forecast the daily average concentration of NO_2, PM_{10}, $PM_{2.5}$, and the maximum average hourly concentration of O_3 levels for the next day, for each station of the air quality monitoring network in Macao. This comprehends six air quality monitoring stations, operated by Macao Meteorological and Geophysical Bureau (SMG), being two of them classified as roadside (Macao Roadside, Ká-Hó Roadside), two as high density residential (Macao Residential, Taipa Residential), and two as ambient background types (Taipa Ambient, Coloane Ambient). Figure 1 represents the air quality monitoring stations spatial location, within the 30 km^2 of Macao region.

Data from 4-year daily series observations, from 2013 to 2016, were used to develop the forecast models, and each of the models was evaluated using 2017 data.

Table 1 Variables used as predictors in the MR and CART models

Variable type		Variable name	Variable description (units)/observations
Air quality		NO_2, PM_{10}, $PM_{2.5}$	Average hourly concentration values (µg/m^3)
		$O_{3\ MAX}$, $CO_{\ MAX}$	Maximum hourly concentration values (µg/m^3)
		16D#, 23D#	23D#: 24-h concentration averaging period between 00H and 23H; 16D#: 24-hour concentration averaging period between 16H of D1 and 15H of D0 eg: PM10_16D1, O3_MAX_23D1
		D0, D1, D2, D3	D0: Forecast Day; D1: Previous Day (Forecast Day-1); D2: Forecast Day-2; D3: Forecast Day-3
Meteo	Upper-air obs.*	H1000, H850, H700, H500	Geopotential height at 1000 hPa, 850 hPa, 700 hPa, 500 hPa (m)/indicator of synoptic-scale weather pattern
		TAR925, TAR850, TAR700	Air temperature at 925 hPa, 850 hPa, 700 hPa (°C)/measure of strength and height of the subsidence inversion
		HR925, HR850, HR700	Relative humidity at 925 hPa, 850 hPa, 700 hPa (%)
		TD925, TD850, TD700	Dew point temperature at 925 hPa, 850 hPa, 700 hPa (°C)
		THI850, THI700, THI500	Thickness at 850 hPa, 700 hPa, 500 hPa (m)/related to the mean temperature in the layer
		STB925, STB850, STB700	Stability at 925 hPa, 850 hPa, 700 hPa (°C)/indicator of atmospheric stability
	Surface obs.	T_AIR_MX, T_AIR_MD, T_AIR_MN	Maximum, average, minimum air temperature (°C)
		HRMX, HRMD, HRMN	Maximum, average, minimum relative humidity (%)
		TD_MD	Average dew point temperature (ground level) (°C)
		RRTT	Precipitation (mm)/associated with atmospheric washout
		VMED	Average wind speed (m/s)/related to dispersion
Other		DD	Duration of the day: number of hours of sun per day (h)
		FF	Week day indicator (flag): weekday = 0, weekend = 1

Meteo, meteorological; *daily sounding at 12H (GMT+8) at King's Park Meteorological Station - Hong Kong Observatory

The first step of the study was to gather a set of meteorological and air quality data, namely (i) meteorological surface observations: hourly observations from automatic weather stations, such as temperature, relative humidity, and dew point temperature collected from the Taipa Grande Meteorological Station; (ii) upper-air observations, such as, geopotential heights, temperature, relative humidity, and dew point temperature at various altitudes, collected from Hong Kong King's Park location; (iii) surface air quality measurements, from SMG's network, of NO_2, PM_{10}, $PM_{2.5}$, and O_3. Other variables were added to the analysis, as the flag for week/weekend day and the daily sunlight period duration. These variables are presented in Table 1.

The next step was to assess data efficiency levels, for each parameter, through the years, in order to reject lower annual efficiencies. The statistical models for Ká-Hó Roadside station were not feasible, due to the lack of sufficient air quality data. Outliers were identified and excluded from the data series. A complimentary analysis was conducted to observe air pollution trends, monthly, weekly, and hourly patterns, and pollution roses.

A preliminary exploratory data analysis, looking at basic statistics, like average, mode, histogram, distribution type, correlation between different variables, and principal component analysis, was performed to identify variables with similar behaviors. This strategy enabled to decide the proper steps to get the best model outcome.

The significance level of 0.05 was used in the linear MR analysis. Some variables initially selected were rejected from the forecast models due to collinearity. The final objective was to obtain prediction models with the lowest possible number of variables but with the maximum explained variance as translated by the R^2. The higher the number of variables used by the model, the higher the risk of compromising the operational forecast, due to lack of information/missing data in case one or more variables are not accessible. SPSS version 25 was used to perform linear MR (stepwise method) and CART analysis.

Model performance was determined recurring to the following parameters: coefficient of determination (R^2) (1), root mean square error (RMSE) (2), mean absolute error (MAE) (3), and Bias (4).

$$R^2 = \frac{\left[\int_{i=1}^{n}\left(f_i - \overline{f}\right) - \left(o_i - \overline{o}\right)\right]^2}{\left[\int_{i=1}^{n}\left(f_i - \overline{f}\right)^2\right]\left[\int_{i=1}^{n}\left(o_i - \overline{o}\right)^2\right]} \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f_i - o_i\right)^2} \tag{2}$$

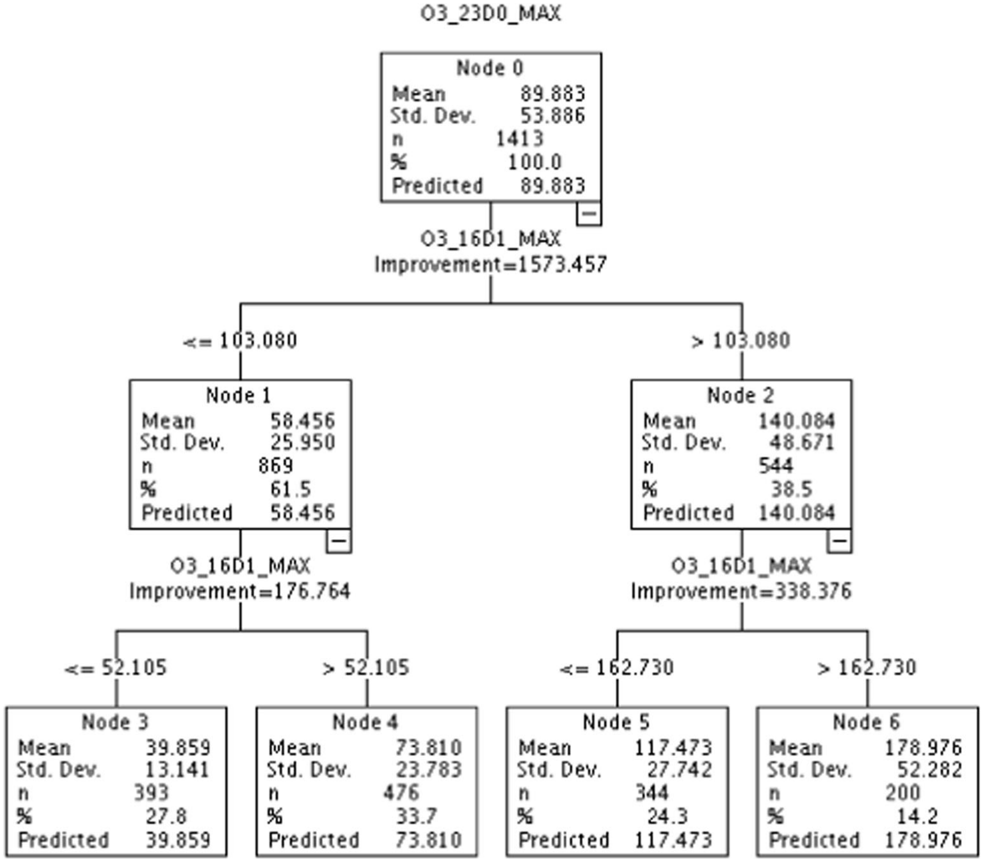

Fig. 2 CART tree obtained for $O_{3\ \mathrm{MAX}}$ prediction at Taipa Ambient station

Table 2 Variables and model equations for each pollutant per air quality monitoring station

Station	Pollutant	Model equations
Macao Roadside	NO_2	$NO_2 = 0.900 \times NO_2\_16D1 + 0.012 \times H850 - 0.168 \times HRMN$
	PM_{10}	$PM_{10} = 0.900 \times PM_{10}\_16D1 + 0.019 \times H850 - 0.270 \times HRMD$
	$PM_{2.5}$	$PM_{2.5} = 0.934 \times PM_{25}\_16D1 + 0.009 \times H850 - 0.128 \times HRMD$
Macao Residential	NO_2	$NO_2 = 0.919 \times NO_2\_16D1 + 0.007 \times H850 - 0.098 \times HRMN$
	PM_{10}	$PM_{10} = 0.884 \times PM_{10}\_16D1 + 0.019 \times H850 - 0.274 \times HRMD$
	$PM_{2.5}$	$PM_{2.5} = 0.915 \times PM_{25}\_16D1 + 0.005 \times H850 - 0.242 \times TD\_MD$
	$O_{3\ MAX}$	$O_{3\ MAX} = 1.123 \times O_{3\_max}\_16D1 - 0.314 \times O_{3\_max}\_23D1 - 0.055 \times HR925 + 0.440 \times T\_AIR\_MX$
Taipa Ambient	NO_2	$NO_2 = 0.915 \times NO_2\_16D1 + 0.004 \times H850 + 0.758 \times STB925$
	PM_{10}	$PM_{10} = 0.891 \times PM_{10}\_16D1 + 0.018 \times H850 - 0.261 \times HRMD$
	$PM_{2.5}$	$PM_{2.5} = 0.918 \times PM_{25}\_16D1 + 0.009 \times H850 - 0.128 \times HRMD$
	$O_{3\ MAX}$	If $[O_{3\ MAX}\_16D1] \leq 103.08$ $O_{3\ MAX} = 1.111 \times O_{3\_max}\_16D1 - 0.207 \times O_{3\_max}\_23D1 - 0.721 \times STB850$ If $[O_{3\ MAX}\_16D1] =]103.08; 162.73]$ $O_{3\ MAX} = 1.237 \times O_{3\_max}\_16D1 - 0.433 \times O_{3\_max}\_23D1 - 1.690 \times STB850$ If $[O_{3\ MAX}\_16D1] > 162.73$ $O_{3\ MAX} = 0.930 \times O_{3\_max}\_16D1 - 0.473 \times O_{3\_max}\_23D1 - 8.608 \times STB850$
Taipa Residential	NO_2	$NO_2 = 0.848 \times NO_2\_16D1 + 0.008 \times H850 - 0.315 \times TDMD$
	PM_{10}	$PM_{10} = 0.894 \times PM_{10}\_16D1 + 0.017 \times H850 - 0.237 \times HRMD$
	$PM_{2.5}$	$PM_{2.5} = 0.937 \times PM_{25}\_16D1 - 0.651 \times TDMD + 0.746 \times TAR925$
	$O_{3\ MAX}$	If $[O_{3\ MAX}\_16D1] \leq 129.05$ $O_{3\ MAX} = 1.043 \times O_{3\_max}\_16D1 - 0.240 \times O_{3\_max}\_23D1 + 0.016 \times H850 - 0.163 \times HRMN$ If $[O_{3\ MAX}\_16D1] =]129.05; 205.47]$ $O_{3\ MAX} = 0.997 \times O_{3\_max}\_16D1 - 0.387 \times O_{3\_max}\_23D1 + 0.055 \times H850 - 0.677 \times HRMN$ If $[O_{3\ MAX}\_16D1] > 205.47$ $O_{3\ MAX} = 1.170 \times O_{3\_max}\_16D1 - 0.482 \times O_{3\_max}\_23D1 + 0.124 \times H850 - 2.632 \times HRMN$
Coloane Ambient	NO_2	$NO_2 = 0.930 \times NO_2\_16D1 - 0.617 \times TDMD + 0.739 \times TAR925$
	PM_{10}	$PM_{10} = 0.875 \times PM_{10}\_16D1 + 0.023 \times H850 - 0.331 \times HRMD$
	$PM_{2.5}$	$PM_{2.5} = 0.903 \times PM_{25}\_16D1 + 0.008 \times H850 - 0.121 \times HRMN$
	$O_{3\ MAX}$	If $[O_{3\ MAX}\_16D1] \leq 113.96$ $O_{3\ MAX} = 1.014 \times O_{3\_max}\_16D1 - 0.197 \times O_{3\_max}\_23D1 + 0.834 \times T\_AIR\_MX - 0.129 \times HRMN$ If $[O_{3\ MAX}\_16D1] =]113.96; 181.61]$ $O_{3\ MAX} = 1.054 \times O_{3\_max}\_16D1 - 0.394 \times O_{3\_max}\_23D1 + 2.676 \times T\_AIR\_MX - 0.597 \times HRMN$ If $[O_{3\ MAX}\_16D1] > 181.61$ $O_{3\ MAX} = 0.666 \times O_{3\_max}\_16D1 - 0.448 \times O_{3\_max}\_23D1 + 7.298 \times T\_AIR\_MX - 1.561 \times HRMN$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - o_i| \tag{3}$$

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i) \tag{4}$$

where f is forecast, $\overline{f}$ is forecast average, o is observation, and $\overline{o}$ is observation average, for each i case to the n number of cases.

Results and discussion

The statistical models based on MR and CART analysis were developed to forecast NO_2, PM_{10}, $PM_{2.5}$, and O_3 concentrations. The final objective is to be able to perform a daily forecast, for the next day, in an operational mode, by running the prediction models after 16H (due to the daily schedules of which the air quality data is made available).

CART analysis was tested mainly in order to better predict the high concentration levels. For NO_2 and PM, CART analysis did not improve the quality of the overall predictions. Therefore, prediction models were based only on one MR model. In the case of O_3 forecast, for three stations (Taipa Ambient, Taipa Residential, and Coloane Ambient), CART analysis allowed to identify split nodes, for which O_3 prediction equations were determined afterwards by using MR for each node. Figure 2 represents an example of the CART trees obtained, in this case for $O_{3\ MAX}$ prediction at Taipa Ambient station.

The output meteorological and air quality variables and equations obtained with MR (or CART and MR, in the $O_{3\ MAX}$ case) are listed in Table 2.

The models were validated with collected data from 2017. The results show a good agreement between modelled and

Table 3 Model performance indicators

Station	Pollutant	Model performance indicator				Model built using only MR or CART and MR	
		R^2	RMSE	MAE	BIAS	MR	CART
Macao Roadside	PM_{10}	0.91	9.2	6.6	1.5	✓	
	$PM_{2.5}$	0.90	5.9	4.0	1.5	✓	
	NO_2	0.89	7.9	5.8	0.9	✓	
Macao Residential	PM_{10}	0.91	8.3	5.8	1.2	✓	
	$PM_{2.5}$	0.86	5.9	3.6	0.9	✓	
	NO_2	0.87	7.8	5.6	-0.2	✓	
	$O_{3\ MAX}$	0.81	23.2	14.0	0.0	✓	
Taipa Ambient	PM_{10}	0.92	6.8	4.5	1.1	✓	
	$PM_{2.5}$	0.89	5.0	3.2	1.1	✓	
	NO_2	0.90	6.1	4.4	0.4	✓	
	$O_{3\ MAX}$	0.82	25.7	15.0	1.3	✓	✓
Taipa Residential	PM_{10}	0.92	6.4	4.1	1.5	✓	
	$PM_{2.5}$	0.89	4.9	3.3	− 0.3	✓	
	NO_2	0.84	6.7	4.6	− 0.5	✓	
	$O_{3\ MAX}$	0.87	21.1	12.2	3.7	✓	✓
Coloane Ambient	PM_{10}	0.93	7.7	5.7	1.9	✓	
	PM_{25}	0.90	5.4	3.6	0.9	✓	
	NO_2	0.85	6.4	4.1	0.0	✓	
	$O_{3\ MAX}$	0.78	27.4	16.9	− 1.5	✓	✓

observed concentrations, being statistically significant at the 95% confidence level. The selected models provide a good relationship between meteorological and air quality variables, when performing an air quality forecast under different situations. Table 3 contains the obtained model performance indicators, such as, R^2, RMSE, MAE, and Bias.

The obtained results performed a better R^2 for PM (between 0.86 and 0.93 and, in all cases, greater for PM_{10} than for $PM_{2.5}$), followed by NO_2 (between 0.84 and 0.90), being the lowest explained variance achieved for O_3 (between 0.78 and 0.87). Models did not show a defined trend on the forecasts by type of station, presenting undistinctive R^2 for roadside, residential, and ambient stations. The monitored and forecasted concentrations, in 2017, for the models with the highest and lowest R^2 are depicted in Figs. 3 and 4, being respectively, the one for PM_{10} Coloane Ambient and $O_{3\ MAX}$ Coloane Ambient, in 2017. The poorest results obtained in Coloane Ambient is related with the fewest cases available to build the model ($N = 546$).

Regarding the RMSE, all models presented the same trend observed for R^2, being the RMSE lower for PM (between 4.9 and 9.2 μg/m^3), followed by NO_2 (between 6.1 and 7.9 μg/m^3), and the highest for O_3 (between 21.1 and 27.4 μg/m^3). In the case of O_3, the high RMSE obtained values were due to

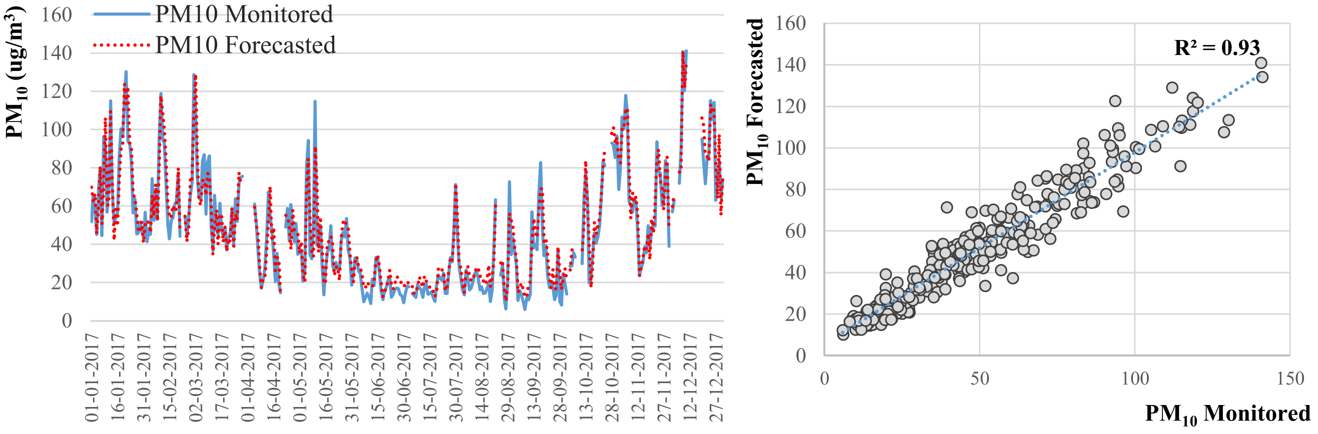

Fig. 3 Observed and predicted PM_{10} concentrations for Coloane Ambient in 2017

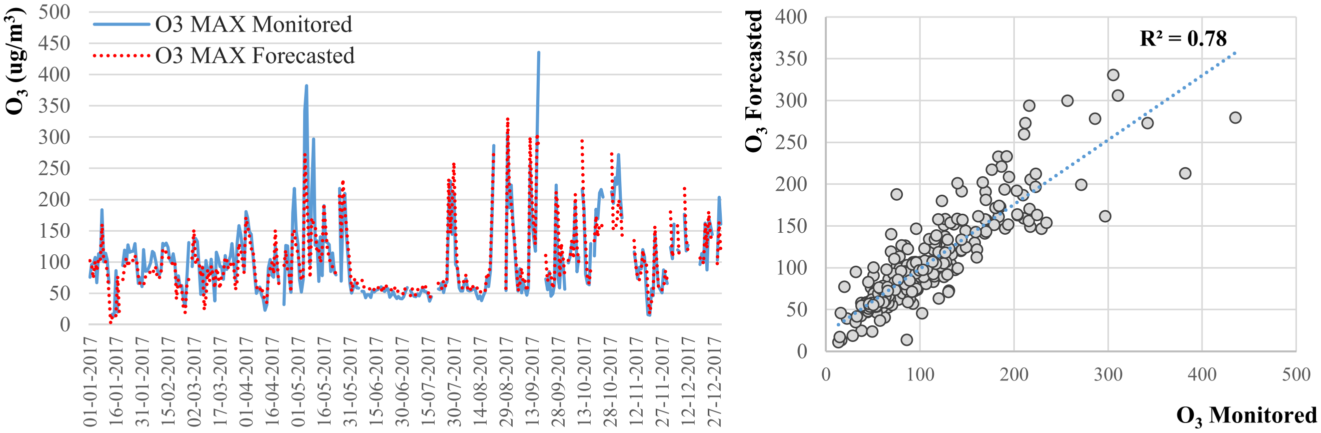

Fig. 4 Observed and predicted $O_{3\ MAX}$ concentrations for Coloane Ambient in 2017

abrupt variations, on consecutive days, influencing the predicted values, since statistical models are sensitive to this kind of fluctuations.

Regarding CART analysis for O_3 prediction, three equation nodes were used. The number of cases considered in each node (N), the coefficient of determination (R^2), the correlation coefficient (r), and the standard error of the estimate are presented in Table 4. The obtained standard error of the estimate, which is a measure of the prediction's accuracy, was higher for higher concentrations prediction categories. The highest obtained standard error of the estimate for node 1 was of 17.2 μg/m^3 in Coloane Ambient station, for node 2 was of 28.8 μg/m^3, and for node 3 was of 43.6 μg/m^3, both in Taipa Residential station. This reflects the difficulty of the model on predicting the highest O_3 concentration ranges. Traffic-related pollutants, such as PM and NO_2, are dependent on meteorological conditions as well as emission rates. Because O_3 is produced in the atmosphere through photochemical processes, the major meteorological factors affecting ozone concentrations are different from those for traffic-related primary pollutants (Choi et al. 2013).

In all the cases, the variable that represents the last 24-h pollutant concentrations (16D1) is the most prevalent, being selected at all the forecast equations (Table 3). The geopotential height at 850 hPa (H_850), indicator of synoptic-scale weather pattern, is also frequently present in the forecast of NO_2 and PM. Specifically, in the case of PM_{10}, relevant variables are H_850 and the medium relative humidity (HRMD), while for $PM_{2.5}$, for both residential stations, average dew point temperature (TD_MD) and air temperature at 925 hPa (TAR_925, a measure of the strength and height of the subsidence inversion) figure in the final equations. Atmospheric stability at 925 hPa and at 850 hPa (STB_925 and STB_850, respectively) figure in final equations in the case of NO_2 and $O_{3\ MAX}$ at Taipa Ambient. This temperature differences between layers provide information about atmospheric stability.

The used statistical methods depend on the past series of data. If the historical data is insufficient, forecasted data will be less reliable. In particular, if emission sources change considerably or if meteorological variables also change due to factors related to new weather patterns eventually motivated by climate change, the data series of the past will not represent the updated situation, and models need to be recalculated with more recent data.

Table 4 CART model performance indicators

Station	Nodes split	N	Model performance indicator		
			R^2	r	Standard error of the estimate
Taipa Ambient	$[O_{3\ MAX}\_16D1] \leq 103.08$	873	0.93	0.97	16.57
	$[O_{3\ MAX}\_16D1] =]103.08; 162.73]$	347	0.97	0.98	22.70
	$[O_{3\ MAX}\_16D1] > 162.73$	200	0.96	0.98	38.59
Taipa Residential	$[O_{3\ MAX}\_16D1] \leq 129.05$	930	0.95	0.98	15.96
	$[O_{3\ MAX}\_16D1] =]129.05; 205.47]$	242	0.97	0.98	28.84
	$[O_{3\ MAX}\_16D1] > 205.47$	99	0.96	0.98	43.62
Coloane Ambient	$[O_{3\ MAX}\_16D1] \leq 113.96$	389	0.94	0.97	17.25
	$[O_{3\ MAX}\_16D1] =]113.96; 181.61]$	106	0.97	0.99	24.32
	$[O_{3\ MAX}\_16D1] > 181.61$	52	0.96	0.98	40.73

Conclusion

The development of statistical models to forecast the daily average concentration of NO_2, PM_{10}, $PM_{2.5}$, and the maximum hourly average concentration of O_3 for the next day, in Macao region, was successfully accomplished for five locations, recurring to MR analysis. In the case of O_3 predictions, CART analysis showed better results, specially improving high concentration levels predictions, assuring a more accurate prediction of critical pollution episodes.

The pollutants for which best results were obtained were PM_{10}, followed by $PM_{2.5}$ and NO_2. The most challenging pollutant forecast was the maximum hourly concentration of O_3, scoring the lowest R^2 (0.78), due to its secondary nature as a pollutant, involved in several atmospheric reactions that depend on the concentrations of other compounds, and also key meteorological conditions, such as sunlight and temperature.

The variables that explained most of the variability, for all pollutants, were the concentration levels measured in the previous 24-h to the operational forecast. For PM and NO_2, the indicator of synoptic-scale weather pattern (geopotential height at 850 hPa parameter), was also a relevant variable.

This work shows that in areas such as Macao, where data may not be easily obtained with a high level of confidence (such as spatially resolved emissions and traffic-related data), this kind of statistical approach becomes an opportunity to obtain a reliable forecast with a clearer understanding of the main factors that affect air quality.

Funding information The work developed was supported by The Macao Meteorological and Geophysical Bureau (SMG). The research work of CENSE is financed by the Fundação para a Ciência e Tecnologia, I.P., Portugal (UID/AMB/04085/2019).

References

Cassmassi JC (1987) Development of an objective ozone forecast model for the South Coast Air Basin. Annual meeting of the Air Pollution Control Association, Conference: 80, Journal Volume: 4, New York, NY (USA), 21-26 Jun Technical Paper 87-71.3; Journal ID: ISSN 0193-9688

Choi W, Paulson SE, Casmassi J, Winer AM (2013) Evaluating meteorological comparability in air quality studies: classification and regression trees for primary pollutants in California's South Coast Air Basin. Atmos Environ 64:150–159. https://doi.org/10.1016/j.atmosenv.2012.09.049

Clapp LJ, Jenkin ME (2001) Analysis of the relationship between ambient levels of O3, NO2 and NO as a function of NOx in the UK. Atmos Environ 35:6391–6405. https://doi.org/10.1016/S1352-2310(01)00378-8

Deng T, Chen Y, Wan Q et al (2018) Comparative evaluation of the impact of GRAPES and MM5 meteorology on CMAQ prediction over Pearl River Delta, China. Particuology 40:88–97. https://doi.org/10.1016/j.partic.2017.10.005

Durão RM, Mendes MT, Pereira MJ (2016) Forecasting O3 levels in industrial area surroundings up to 24 h in advance, combining classification trees and MLP models. Atmos Pollut Res 7:961–970

Entwistle MR, Gharibi H, Tavallali P et al (2019) Ozone pollution and asthma emergency department visits in Fresno, CA, USA, during the warm season (June–September) of the years 2005 to 2015: a time-stratified case-crossover analysis. Air Qual Atmos Heal 12: 661–672. https://doi.org/10.1007/s11869-019-00685-w

He D, Zhou Z, He K et al (2000) Assessment of traffic related air pollution in urban areas of Macao. J Environ Sci 12:39–46

Kumar R, Barth MC, Pfister GG et al (2018) How will air quality change in South Asia by 2050? J Geophys Res Atmos 123:1840–1864. https://doi.org/10.1002/2017JD027357

Lee M, Brauer M, Wong P et al (2017) Land use regression modelling of air pollution in high density high rise cities: a case study in Hong Kong. Sci Total Environ 592:306–315. https://doi.org/10.1016/j.scitotenv.2017.03.094

Liu JC, Peng RD (2018) Health effect of mixtures of ozone, nitrogen dioxide, and fine particulates in 85 US counties. Air Qual Atmos Heal 11:311–324. https://doi.org/10.1007/s11869-017-0544-2

Lopes D, Hoi KI, Mok KM et al (2016) Air quality in the main cities of the pearl river delta region. Glob Nest J 18:794–802

Martinez NM, Montes LM, Mura I, Franco JF (2018) Machine Learning Techniques for PM 10 Levels Forecast in Bogotá. In: 2018 ICAI Workshops (ICAIW). IEEE, pp 1–7. doi: https://doi.org/10.1109/ICAIW.2018.8554995

Oduro SD, Ha QP, Duc H (2016) Vehicular emissions prediction with CART-BMARS hybrid models. Transp Res Part D Transp Environ 49:188–202. https://doi.org/10.1016/j.trd.2016.09.012

Reid N, Yap D, Bloxam R (2008) The potential role of background ozone on current and emerging air issues: an overview. Air Qual Atmos Heal 1:19–29. https://doi.org/10.1007/s11869-008-0005-z

Sheng N, Tang UW (2013) Risk assessment of traffic-related air pollution in a world heritage city. Int J Environ Sci Technol 10:11–18. https://doi.org/10.1007/s13762-012-0030-1

SMG (2014) Climate in Macao. SMG/ Macao Meteorological and Geophysical Bureau. Available at: http://www.smg.gov.mo/smg/climate/e_climaintro.htm. Accessed 1 June 2019

SMG (2019) Annual summary of air quality in Macao – 2018. SMG/ Macao Meteorological and Geophysical Bureau. Available at: http://www.smg.gov.mo/smg/airQuality/pdf/IQA_2018_PT.pdf. Accessed 1 June 2019

Tong CHM, Yim SHL, Rothenberg D et al (2018a) Assessing the impacts of seasonal and vertical atmospheric conditions on air quality over the Pearl River Delta region. Atmos Environ 180:69–78. https://doi.org/10.1016/j.atmosenv.2018.02.039

Tong CHM, Yim SHL, Rothenberg D et al (2018b) Projecting the impacts of atmospheric conditions under climate change on air quality over the Pearl River Delta region. Atmos Environ 193:79–87. https://doi.org/10.1016/j.atmosenv.2018.08.053

US EPA (2003) Guidelines for Developing an Air Quality (Ozone and PM2.5) Forecasting Program. doi: EPA-456/R-03-002. Available at: https://nepis.epa.gov/Exe/ZyPURL.cgi?Dockey=2000F0ZT.TXT. Accessed 1 June 2019

WHO (2018) Ambient (outdoor) air quality and health. https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health. Accessed 2 Jul 2019

WHO (2019) Air pollution and health: summary. https://www.who.int/airpollution/ambient/about/en/. Accessed 2 Jul 2019

Wiśniewska K, Lewandowska AU, Staniszewska M (2019) Air quality at two stations (Gdynia and Rumia) located in the region of Gulf of Gdansk during periods of intensive smog in Poland. Air Qual Atmos Heal 12:879–890. https://doi.org/10.1007/s11869-019-00708-6

Xie J, Liao Z, Fang X et al (2019) The characteristics of hourly wind field and its impacts on air quality in the Pearl River Delta region during

2013–2017. Atmos Res 227:112–124. https://doi.org/10.1016/j.atmosres.2019.04.023

Zhang J, Ding W (2017) Prediction of air pollutants concentration based on an extreme learning machine: the case of Hong Kong. Int J Environ Res Public Health 14:1–19. https://doi.org/10.3390/ijerph14020114

Zheng J, Zhang L, Che W et al (2009) A highly resolved temporal and spatial air pollutant emission inventory for the Pearl River Delta region , China and its uncertainty assessment. Atmos Environ 43:5112–5122. https://doi.org/10.1016/j.atmosenv.2009.04.060

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.