The role of gap-filling observational data in air quality data fusion methods: a case study with CALIOPE PM2.5

Author: Ada Barrantes Cepas[1]
[1]*Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain.*

Advisors: Cristina Carnerero[2], Jan Mateu Armengol[2,3], Mireia Udina[1]
[2]*Barcelona Supercomputing Center, Earth Sciences Department, Barcelona, Spain*
[3]*Department of Fluid Mechanics, Universitat Politècnica de Catalunya, Barcelona, 08034, Spain*

Reliable air quality data are vital for informed decision-making, enabling evidence-based mitigation strategies to improve public health and sustainability. Data-fusion methods combining physics-based air quality models with observational data provide reliable results with full spatial coverage. This study quantifies the impact of imputing missing observational data in these data-fusion methods. We focus on PM2.5 for the Catalonia region during 2019, for which data availability is strongly limited. We first present straightforward gap-filling methodologies, such as linear interpolation and persistence. We then compare these techniques with a state-of-the-art artificial intelligence gap-filling method based on the Gradient Boosting Machine algorithm trained with several years of data (2019, 2021, 2022). To assess gap-filling methodologies, we generate random gaps of varying characteristics identifying the optimal technique for each gap size and availability. Finally, we study how these methods affect the data fusion process applied to the mesoscale air quality model CALIOPE. The PM2.5 output of this system has a horizontal spatial resolution of 1 km x 1 km on a daily scale. The data fusion method uses universal kriging, a geostatistical technique based on a regression model and the spatial correlation between the model and observational data. Data fusion results significantly improve from the raw model estimations, with +24 % and +61 % for the r-value, not using gap-filling of observational data and using it, respectively. Notably, the method's effectiveness depends on the availability of observations, performing better with GBM-filled data.

I. INTRODUCTION

Air pollution is the foremost environmental health problem in the European Union (EU) (*WHO* 2021). Air quality has emerged as a pressing concern of pollution's impact on public health, ecosystems, and the economy. Fine particulate matter (PM2.5) is particularly harmful, causing over 300,000 premature deaths annually in Europe (*Commission* 2024). Directive 2008/50/EC (*EC* 2008) on ambient air quality and cleaner air for Europe introduced specific objectives targeting the reduction of population exposure to PM2.5, aiming for an annual average concentration lower than 25 µg/m^3.

PM2.5 consists of tiny particles with 2.5 micrometers of diameter or less, which can penetrate deep into the lungs and even enter the bloodstream. This pollutant is a mixture of solid particles and liquid droplets, originating from various sources such as vehicle emissions, industrial processes, residential heating, and natural sources like wildfires or dust. The health effects of PM2.5 are well-documented (*Xing et al.* 2016), including respiratory and cardiovascular diseases, lung cancer, and adverse birth outcomes. Chronic exposure to PM2.5 is linked to reduced life expectancy and increased mortality rates.

In addition to its health impacts, PM2.5 also affects the environment by contributing to the formation of smog and acid rain, which can harm wildlife, damage forests, and degrade water quality. Economically, the burden of air pollution manifests through healthcare costs, reduced labor productivity, and loss of biodiversity, which can impact tourism and agriculture.

Monitoring stations are essential for assessing air quality. However, they have limited spatial representativeness, leaving large extensions of areas without appropriate observational data. Conversely, numerical air quality systems provide comprehensive spatial coverage. Modeled data are affected by persistent uncertainties, mainly due to emission inventory inaccuracies and the complexity of atmospheric processes involved in pollution transport. Data fusion methods offer bias-corrected air quality maps with full spatial coverage (*Horálek* 2006). Nonetheless, there is a strong dependence on observational data availability to ensure reliable results of data fusion methods.

The importance of this study relies on the demand for precise and in-time pollution prediction information in regions lacking air quality monitoring stations. Moreover, assessing cities and areas with known elevated pollution values is relevant for implementing effective control measures and initiatives to reduce pollution.

This work aims to improve the outputs of CALIOPE, a regional air quality modeling system tailored to the northeast region of Spain, namely Catalonia. To this end, we use a data fusion method to combine observational and modeled data. CALIOPE system operationally provides air quality forecasts at 24 h and 48 h. We post-process 2019 daily outputs for PM2.5 in Catalonia with a spatial resolution of 1 km x 1 km (*Baldasano et al.* 2011). Another objective of this study is to analyze the improvement achieved by applying artificial intelligence (AI) techniques to fill data gaps compared to simpler methods. Specifically, we focus on the Gradient

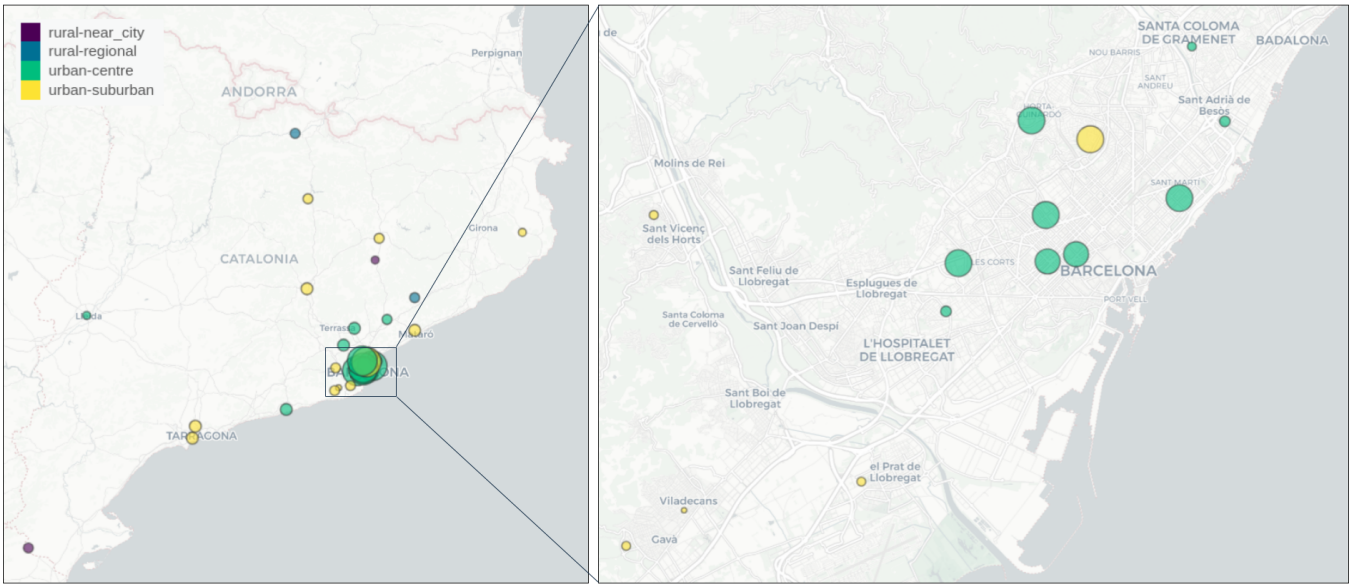

FIG. 1. Domain of study and PM2.5 monitoring stations from XVPCA. The circle size represents the total data availability (%) during 2019, while the color indicates the station category (Table III. Appendix).

Boosting Machine (GBM), a machine learning technique proven highly effective in imputing missing air quality data (*Su* 2020). We also evaluate the influence of each proposed predictor on the imputation process, given that GBM employs various predictors to estimate missing values. This comprehensive assessment aims to highlight the advantages of AI-based gap-filling methods over traditional approaches and to understand the contribution of each predictor in enhancing the accuracy and reliability of air quality data. We demonstrate the importance of pre-processing raw observational data to correct air quality information.

II. METHODOLOGY & DATA ANALYSIS

The correction of the modeling output for PM2.5 was performed using a data fusion approach, both with and without the implementation of gap-filling techniques. Before this, we conducted a benchmark to evaluate the performance of various gap-filling methods.

A. Study domain and observational PM2.5 data

Daily PM2.5 observational data for 2019 are obtained from the Catalan Air Pollution Monitoring and Forecasting Network (XVPCA) stations. There are 30 measurement stations in Catalonia's region, with an average daily data availability of over 55 % (Fig. 1). Of these, 14 are urban-center traffic monitoring stations, 12 are urban-suburban stations, 2 are rural-regional, and the remaining two are rural near-city stations.

The study domain is Catalonia, located in the northeast of Spain, covering an area of 32,107 km^2 with a population of over 7.901 million on 1st of January 2023 (*Institut_d'Estadística_de_Catalunya* 2024). Although there are 947 municipalities across the region, 95 % of the population resides in only 300 of them, considered urban areas. The heterogeneous terrain, varied land use, and diverse vegetation contribute to unique local conditions and challenging pollutant prediction.

The orography of Catalonia can be classified into three main areas: the central depression, the coastal border delineated by the Prelitoral and Litoral mountain ranges, and the Pyrenees and Pre-Pyrenees region. Air pollution typically accumulates and is dispersed by winds in the central depression, although it can occasionally become trapped due to the surrounding mountain ranges. Local climatic features and land-sea breezes significantly impact the dispersion of pollutants, particularly during the summer.

Regarding PM2.5 composition, marine sources contribute less than 1 %, crustal sources contribute 8 %, and anthropogenic sources contribute 73 % (*Querol et al.* 2001). The principal PM2.5 anthropogenic sources in Catalonia originate from its industrial network and road traffic along the Metropolitan Area of Barcelona (AMB), the central depression, and the province of Tarragona.

B. Air quality model and data fusion methodology

CALIOPE is an air quality prediction modeling system (*Baldasano et al.* 2011) that integrates the meteorological model WRF-ARW (Advanced Research Weather Research and Forecasting) (*Skamarock and Klemp* 2008),

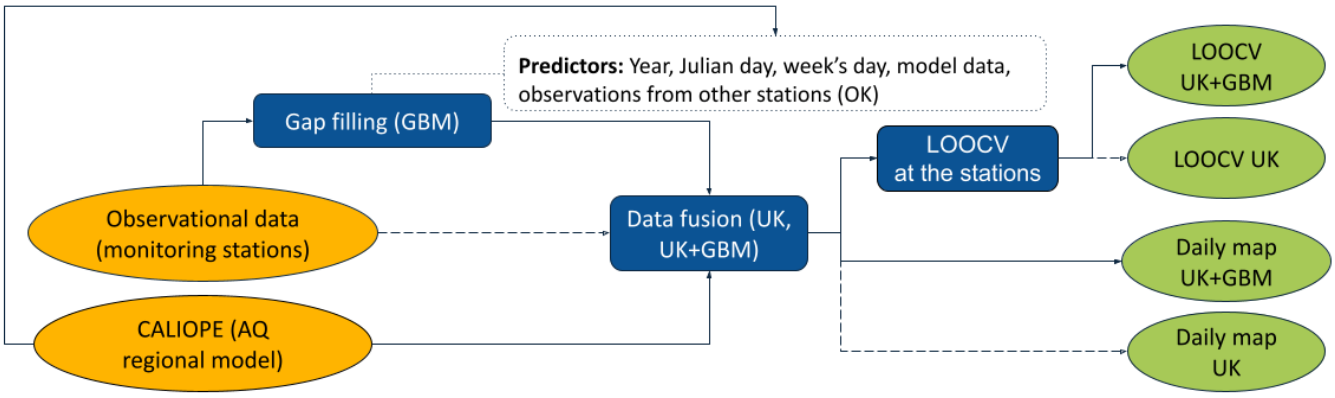

FIG. 2. Workflow of the data fusion methodology. Datasets are represented as circular elements, with orange indicating inputs and green indicating outputs. Squared solid boxes denote processes, while dashed white boxes represent supplemental information considered for these processes. Solid arrows illustrate the steps involved in correcting the model using GBM, while dashed arrows indicate the correction of the model without gap filling.

the chemical transport model CMAQ (Community Multi-scale Air Quality modeling system) (*Byun and Schere* 2006), the emission model HERMESv3 (*Guevara et al.* 2019), and the mineral dust atmospheric model BSC-DREAM8b (*Nickovic et al.* 2001, *Pérez et al.* 2006). The mother domain of CALIOPE runs for Europe at a 12 km x 12 km spatial scale, the nested domains are consecutively the Iberian Peninsula (4 km x 4 km) and Catalonia (1 km x 1 km). The model is the foundation data to be corrected based on the XVPCA observations.

The data fusion methodology illustrated in Fig. 2 is employed to improve model performance, particularly using Ordinary Kriging (OK) and Universal Kriging (UK) (*Hengl and Rossiter* 2007, *Horálek* 2006). We integrate the observational data with the modeled data to conduct data fusion and generate the final corrected model maps. In one case, we fill the gaps in the observational data. In the other case, we maintain the gaps to evaluate the gap-filling role. Following the data fusion process, we employ Leave-One-Out Cross-Validation (LOOCV) to validate the results at each station with the values obtained from other stations. Numerous studies have showcased promising outcomes by applying these techniques to air pollution modeling (*Huang* 2018, *Lin et al.* 2020).

Ordinary Kriging is a spatial interpolation method that estimates values at specific locations based on nearby data points (*Pardo-Iguzquiza and Chica-Olmo* 2008). The estimation involves multiplying each observed data point by its corresponding weight and summing them together. The interpolation model is fitted using a semivariance function, which measures the spatial correlation between two locations as a function of their distance. The objective is to identify a theoretical model that closely aligns with the spatial semivariance structure observed in the data. In our case, we adjust a semivariance function for each day following the Stein theoretical model (*Stein* 1986), aiming to represent PM2.5 spatial variability accurately.

Universal Kriging (UK) is a geostatistical method utilized to estimate unknown values in geographical fields while providing estimates of their variances (*Cressie* 1993). In this approach, the corrected data is predicted through a combination of elements including the linear regression function, the spatially correlated stochastic variation, and the intrinsic noise of the geographical space (residuals). Initially, a linear regression is performed between the observed values and the model's raw data. Then, the residual values are calculated at each monitoring station, representing the difference of each observed value from the trend line. Subsequently, a variogram of the residual values is constructed assuming a spatial correlation of residuals. Finally, the linear regression is applied to all points and then adjusted using the residual value interpolated using ordinary kriging. UK follows the relation:

$$Z(\mathbf{x}) = f(\mathbf{x}) + e(\mathbf{x}) = \sum_{l=0}^{L} a_l f_l(\mathbf{x}) + \sum_{l=0}^{L} b_l e_l(\mathbf{x}) \qquad (1)$$

where Z is the predicted value at the target point $\mathbf{x}$ (2-dimensional), $f(\mathbf{x})$ is the linear regression model and is applied at each point by its estimated coefficients a_l. Finally, $e(\mathbf{x})$ is the residual function, where b_l represents the ordinary kriging weights determined by the spatial dependence structure of the residuals. L corresponds to the total number of covariables, in our case, there is just one which is the CALIOPE raw model.

The Leave-One-Out Cross-Validation (LOOCV) methodology can be used to evaluate data fusion skills (*Le Rest et al.* 2014). It evaluates the performance at each station through iterative processes. During each round, data from one monitoring station is excluded, and the remaining stations' data are used to estimate the value at the precise station. Subsequently, statistical values are computed by comparing the interpolated value with the actual observed value that was previously excluded.

For the LOOCV results, we present the mean bias (MB), the root mean square error (RMSE), the correlation coefficient (r), and the coefficient of efficiency (COE), defined as follows:

$$MB = \frac{1}{N}\sum_{i=1}^{N}(M_i - O_i) \tag{2}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(M_i - O_i)^2} \tag{3}$$

$$r = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{M_i - M}{\sigma_M}\right)\left(\frac{O_i - O}{\sigma_O}\right) \tag{4}$$

$$COE = 1 - \frac{\sum_{i=1}^{N}|M_i - O_i|}{\sum_{i=1}^{N}|O_i - O|} \tag{5}$$

where N is the total number of observations, O_i and M_i are the observed and modeled i values, their means are O and M, and their standard deviation are σ_O and σ_M, respectively. We aim for COE and r values close to 1, a MB close to zero, and a small RMSE to indicate accurate and reliable model predictions.

C. Gap-filling methodology

The availability of PM2.5 air quality data at monitoring stations is occasionally restricted, resulting in numerous days with missing values throughout the year. Hence, there is an interest in filling these data gaps by applying machine learning techniques, such as gradient-boosting machine algorithms or other straightforward gap-filling methods.

To ensure a fair comparison, we selected two extreme study cases. The first is the urban-center Eixample's station with 87 % observational availability in 2019. This station is located in Barcelona and has many nearby stations, which aids the GBM's predictions. As further explained in Section II.C.2, this situation implies that GBM has more values to generate the predictor of the interpolated value from other stations done with ordinary kriging. In contrast, the second selected is the rural-near-city La Sènia station, in Montsià, near the Catalan border with the Valencian Community. This station has limited observational data (45 % availability) and is isolated from other stations, which is expected to decrease the performance of the GBM.

1. *Straight-forward benchmarking methods*

We conducted a comparative analysis between two simple gap-filling techniques and the GBM method. This comparison allows us to contextualize the performance of GBM and evaluate whether its increased complexity is justified. This benchmark involves persistence and linear interpolation techniques. Persistence entails repeating the previous day's value, while linear interpolation performs a linear regression between the values at the boundaries of the gap.

We examine these proposed techniques in the two selected stations, alongside the GBM, to evaluate examples of extreme cases. Therefore, from all the available data, we selected a percentage of it as training data (train fraction) and utilized the remaining values as test data (gaps).

We analyze the influence of different gap sizes and data availability on the gap-filling results. Figure 3 shows the occurrences of gap sizes across all stations in Catalonia for 2019. Although the most frequent gap size is one day, we analyzed sizes from 1 to 7 days. The stations' availability does not follow a regular distribution, with some stations exhibiting high availability while others have limited availability (Fig. 1). To assess the impact of availabilities in gap-filling techniques, we generated random gaps of constant size, varying the percentage of data used as training from 30 % to 90 %. However, there is an intrinsic limit of maximum possible not adjacent gaps; thus, the availability is restricted. For instance, when we do gaps with size 1, we cannot have an availability (train fraction) smaller than 50 %.

As performance results may depend on the days we considered as gaps, we calculated the mean statistical values over 100 gap distributions, generating gaps with the same size and frequency but at different positions. The statistical values used to verify the benchmark are the coefficient of correlation (r), the mean bias (MB), and the root mean squared error (RMSE).

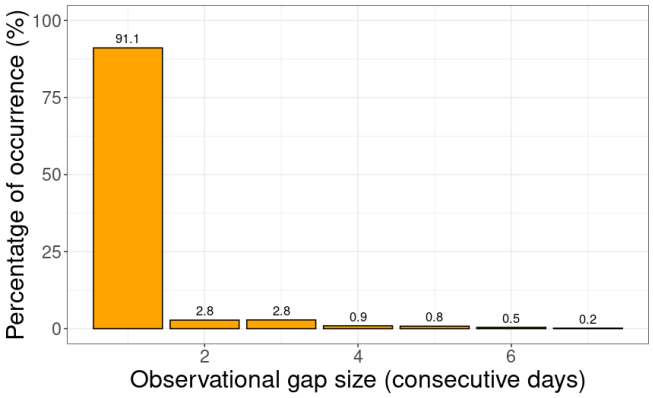

FIG. 3. Histogram distribution of consecutive days without data (gap size) occurrence throughout all the stations in 2019.

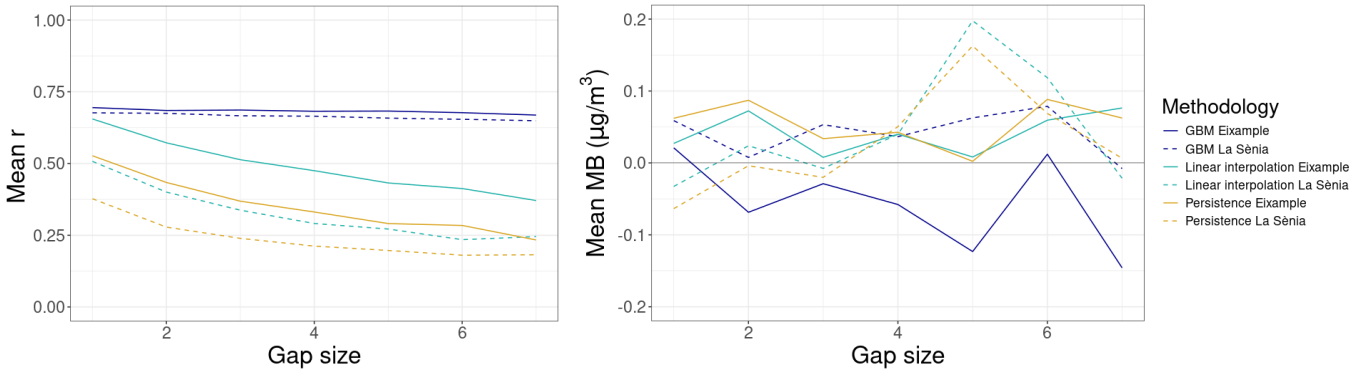

FIG. 4. Gap-filling techniques performance in Eixample's (solid lines) and La Sènia's (dashed lines) stations over different gap sizes. The results comprehend the mean values of the selected range of train fractions (from 30 % to 90 %). GBM performance is represented in dark blue, linear interpolation in sea blue, and persistence in golden. **Left:** Mean correlation coefficient for all train fractions. **Right:** Mean bias (µg/m^3) for each technique.

2. *Machine learning based technique*

GBM is an artificial intelligence method based on boosting, where several simple and ineffective prediction models are combined to produce a more effective overall model. At each iteration, it builds an ensemble of decision trees, with every tree correcting the errors made by the previous trees. The final prediction is formed by combining the predictions of each tree in the ensemble. We train the GBM model to predict the target variable of the daily average PM2.5 based on a selection of predictors. The chosen predictors are the day of the year (Julian Day), the weekday, the year, the modeled raw data from CALIOPE, and the interpolated observational data from all other stations. Meteorology or the synoptic state of the atmosphere is not included as a predictor, as it is already integrated into the CALIOPE modeled data.

The machine learning (GBM) algorithm implemented is in the R package GBM (*Greenwell and Developers* 2022), and it has shown better accuracy in the results when compared to other learning algorithms (*Caruana and Niculescu-Mizil* 2006). The use of the GBM model entails a selection of hyperparameters. We have manually searched for a set of them that produce acceptable results. However, a formal optimization of the GBM's hyperparameters is required and we plan to perform it as future work. The selected hyperparameters are the following: 500 trees, an interaction depth of 1, a shrinkage rate of 0.01, and 5 cross-validation folds. The computations are executed on a single core.

Our case study focuses on 2019 due to the 100 % availability of CALIOPE-modeled data for each day of the year. Given that GBM needs a substantial dataset for training, we opted to utilize data from 2019, 2021, and 2022 to increment its performance. We excluded 2020 due to irregularities in the model and the observational datasets. For instance, 2020 had an atypical PM2.5 pattern due to mobility restrictions (*Querol et al.* 2021), which may introduce additional noise to the GBM, reducing its effectiveness.

III. RESULTS

The results are categorized into three main sections. The first section examines the performance of the gap-filling techniques proposed in this study. Subsequently, we implemented the most effective gap-filling method on our dataset and proceeded with the data fusion process. However, the data fusion is also conducted without filling the gaps to evaluate its impact on the results. Finally, the last section presents the annual concentration values obtained within our study domain and evaluates them under the current air quality legislation.

A. Analysis of gap-filling techniques

Figure 4 illustrates the performance of each gap-filling technique as a function of the gap-sized averaged over all proposed train fractions. A noticeable decline in the correlation coefficient is observed for both linear interpolation and persistence as the gap size increases, applicable to both stations. In contrast, the GBM correlation coefficient remains relatively constant across varying gap sizes, and the difference between stations is less pronounced than the other techniques. La Sènia exhibits a lower correlation coefficient due to its limited data availability and isolated conditions, which primarily impacts the predictor of the interpolated value from other stations in GBM's model.

Both linear interpolation and persistence techniques tend to slightly overestimate the data, except for GBM at Eixample's station, which vaguely underestimates the values. The performance of these techniques is more similar for smaller gap sizes and diverges when the gap size exceeds four days.

Small dependence is noticed over train fraction for

TABLE I. Benchmark's statistical results with a reference gap size of 1 day and 60 % train fraction, in Eixample's and La Sènia's stations. The mean value and standard deviation over 100 different gap distributions are presented for the correlation coefficient, mean bias (µg m^{-3}), and root mean squared error (µg m^{-3}) for the GBM, linear interpolation, and persistence gap-filling techniques.

		Correlation coefficient		**Mean Bias** (µg m^{-3})		**RMSE** (µg m^{-3})	
Station	**Methodology**	$\overline{r}$	σ_r	$\overline{MB}$	σ_{MB}	$\overline{RMSE}$	σ_{RMSE}
Eixample	GBM	0.69	0.05	-0.01	0.48	5.1	0.58
	Linear interpolation	0.65	0.03	0.06	0.46	5.5	0.32
	Persistence	0.52	0.07	0.09	0.53	6.9	0.51
La Sènia	GBM	0.67	0.05	0.05	0.39	3.6	1.56
	Linear interpolation	0.51	0.05	0.01	0.30	2.4	0.17
	Persistence	0.36	0.08	-0.02	0.30	3.0	0.20

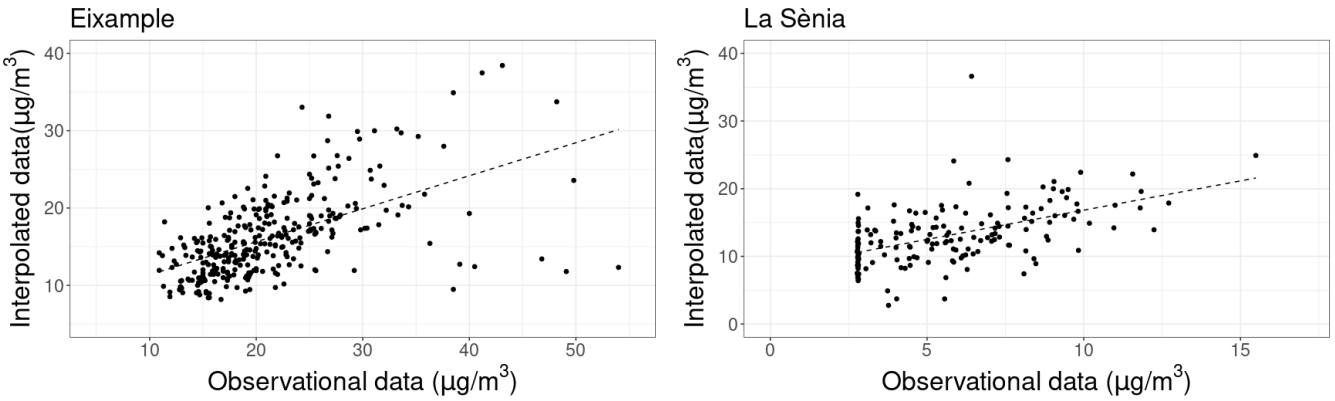

FIG. 5. Dependence of interpolated data from other stations at a precise station using ordinary kriging with the observational data. The black dashed line represents the linear regression. **Left:** Eixample's station with 0.57 correlation coefficient and -5.08 µg m^{-3} mean bias. **Right:** La Sènia's station with 0.25 correlation coefficient and 7.00 µg m^{-3} mean bias.

both techniques, except GBM, which shows a slight increase in the correlation coefficient around the 80 % train fraction (Figure not shown).

In Table I, we present the numerical statistical results of the benchmark, considering a gap size of one day and a train fraction of 60 %, which reflects the average real case in our dataset. The correlation coefficient is consistently higher for GBM at both stations. The mean bias (MB) and root mean squared error (RMSE) are also better for GBM at the Eixample station. Moreover, at La Sènia, linear interpolation exhibits lower absolute MB and RMSE values. This indicates that while GBM is the most effective technique for non-isolated stations, it also performs well in isolated stations.

However, there is a slight decrease in GBM's performance for La Sènia's station. We attribute this decrease to the lack of nearby stations, which affects the predictor of interpolated observations from other stations. To illustrate the quality of this predictor, Fig. 5 shows the correlation between observations and interpolated values from nearby stations. As expected, observations from the Eixample's station better correlate with a 0.57 correlation coefficient, compared to the 0.25 correlation coefficient from La Sènia.

B. Data fusion results

Once having established GBM as the most efficient technique for gap-filling, we proceeded with the data fusion process. We conducted data fusion using the observational dataset, filling the gaps with (UK+GBM) and without (UK). Subsequently, we analyzed the results at each station in LOOCV and across the entire study domain.

1. LOOCV results at the stations

Figure 6 shows the difference in the squared correlation coefficient between the post-processing LOOCV datasets and the actual raw estimations of CALIOPE's model at each station. The main discrepancy is observed in the Metropolitan Area of Barcelona when comparing the post-processed UK with the raw model (Fig. 6a). When gaps are filled and we compare the post-processed UK+GBM with the raw model, differences are evident across all regions. Filling the gaps improves data availability from stations outside the main Barcelona region. This improvement is most pronounced between

TABLE II. Data fusion results comparing the observational data at the stations with the data from the raw model, the post-processing with only UK, and the post-processing with GBM and UK in LOOCV. The statistical parameters are, from left to right: the mean bias (µg m^{-3}), the root mean squared error (µg m^{-3}), the correlation coefficient, and the coefficient of efficiency.

	MB (µg m^{-3})	**RMSE** (µg m^{-3})	**r**	**COE**
CALIOPE's raw model	-9.00	11.06	0.46	-0.66
Correction with UK	0.42	6.81	0.57	0.29
Correction with GBM & UK	-0.10	4.88	0.74	0.39

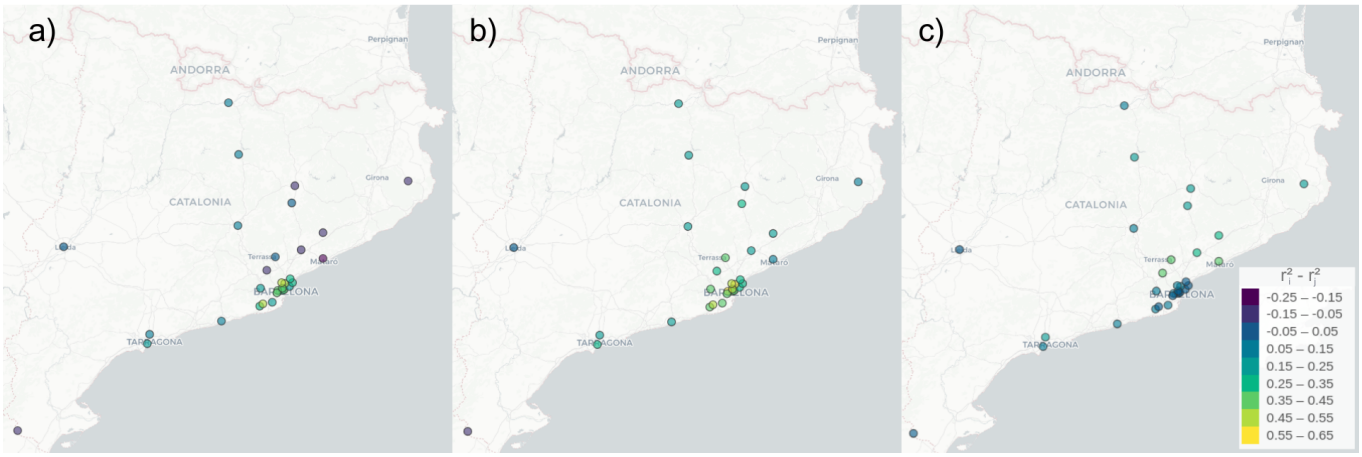

FIG. 6. Squared correlation coefficient (r^2) anomaly in LOOCV between **a)** UK (i) and the raw model (j), **b)** UK+GBM (i) and the raw model (j), and **c)** UK+GBM (i) and UK (j).

UK+GBM and UK in the outskirts of Barcelona, where data availability is small as depicted in Figure 1.

Furthermore, we calculate the relative influence of GBM's predictors in each station (Fig. 7). This influence has been computed based on the methodology proposed by Friedman (*Friedman* 2001), in which the relative importance of each predictor is associated with the reduction in the GBM cost function. The predictor with the highest impact on GBM performance is the value interpolated from other observations at the station. None of the other predictors exceed 25 % influence on the results. The Julian day of the year is the second most significant parameter, suggesting a link between seasonal changes and local climatic conditions throughout the years.

The station with the smallest influence from other observations and the largest influence from the model's data corresponds to La Sènia. As observed in Figure 5, there is a weak correlation between interpolated values and measurements at this station, indicating that this predictor may be less effective in GBM adjusted for La Sènia, and will require more support from the model's data.

Once LOOCV is performed at the stations, we can compare the concentrations obtained through data fusion or the raw model with the actual measurements (Table II). The correlation coefficient increases substantially with data fusion, indicating improved consistency with observed values. It increases by up to 24 % with data fusion using UK alone and by 61 % with UK+GBM

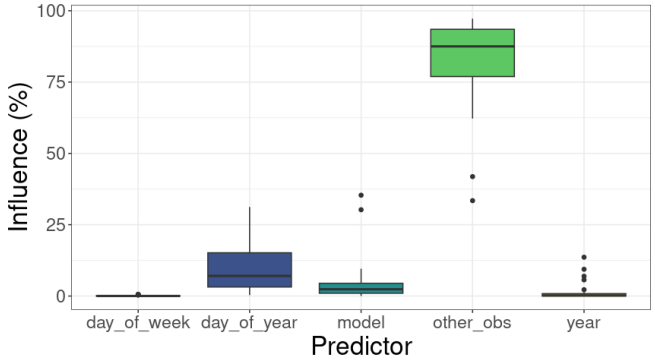

FIG. 7. Influence of GBM's predictors on its performance across all the stations. The box represents the first, second (median), and third quartiles, corresponding to the three consecutive lines. The lower and upper extreme values display the minimum and maximum of the distribution, while any outliers are marked as single values below or above these extremes.

compared to the raw model. The mean bias (MB) decreases considerably in absolute value after data fusion: -95.3 % and -98.8 % without and with gap-filling respectively. Moreover, while the model initially tended to underestimate the observational measurements, data fusion with GBM still leads to underestimation, whereas the data fusion without gap-filling mostly overestimates. The

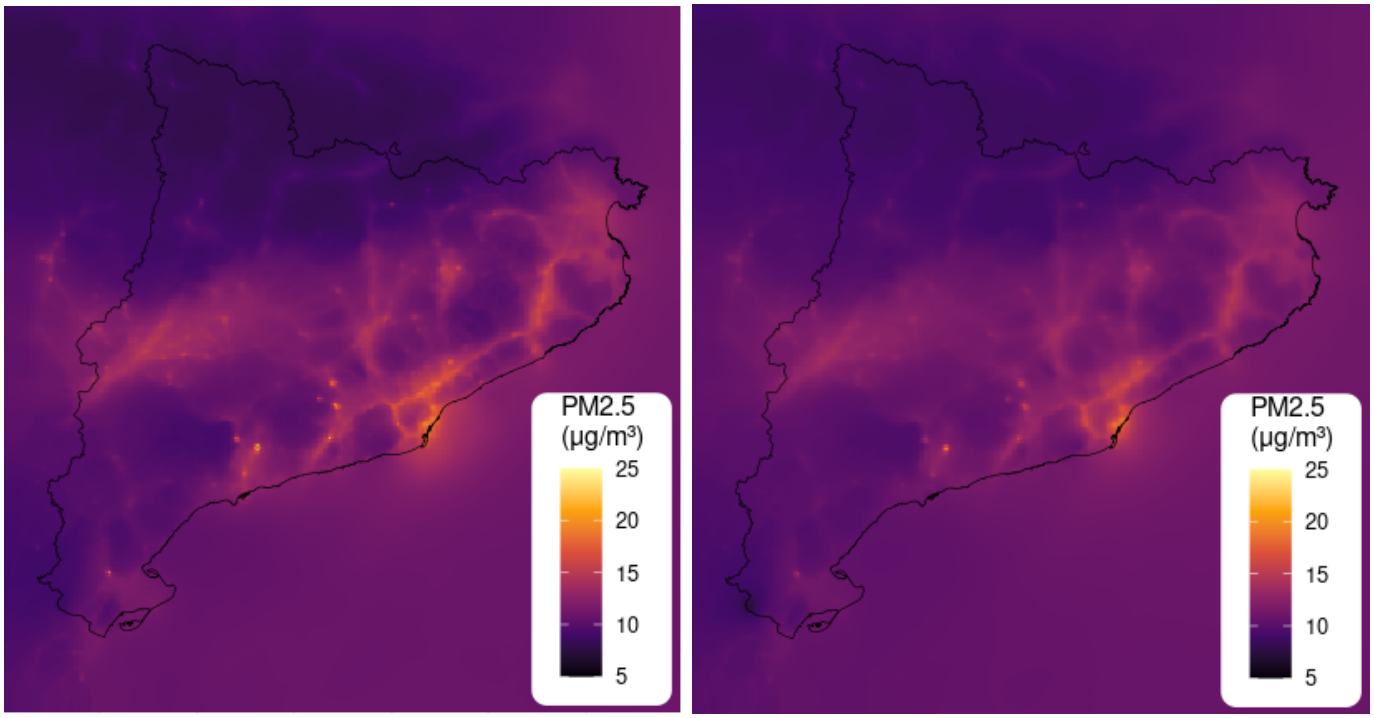

FIG. 8. Annual mean concentrations (µg m^{-3}) in Catalonia's domain. **Left:** Correction only with the UK. **Right:** Correction with UK+GBM.

RMSE also decreases, reflecting higher accuracy. Finally, the coefficient of efficiency and the index of agreement indicate better results when applying GBM before data fusion. All these metrics consistently highlight the importance of gap-filling methodologies before data-fusion methods.

2. *Annual limits PM2.5 evaluation*

Data fusion enables the correction of the model across its entire grid. Figure 8 shows the 2019 annual mean concentration of PM2.5 using UK and UK+GBM. The data fusion with only the UK reaches higher concentrations (Fig. 8 left), indicated by lighter colors. The highest values are distributed around the major roads and the main cities of Catalonia. There is an isolated high value around Valls (41°17'18"N 1°15'03"E), where the concentration is not as high as with UK+GBM data fusion. Valls has significant industrial activity, including factories and manufacturing plants, that emits fine particulate matter and other pollutants. The AMB and its surroundings have high concentrations of PM2.5 due to the combination of industrial emissions, high traffic volumes, population density, port activities, and local meteorological conditions.

The weekly evolution of concentrations across all stations (Fig. 9) demonstrates significant improvement through data fusion. The model consistently underestimates concentrations but consistently tracks the overall

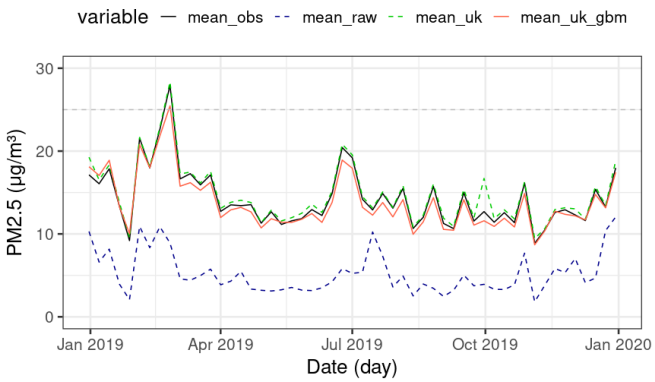

FIG. 9. Temporal variation of mean weekly PM2.5 concentration (µg m^{-3}) of mean stations. The gray dashed line indicates the mean annual limit legislated.

trend. Minimal differences are seen between data fusion with and without gap-filling. GBM tends to underestimate values but accurately predicts variations with high confidence. Conversely, not doing gap-filling may result in occasional artificial spikes, as observed at the beginning of October 2019.

The largest concentrations are seen at the beginning of the year, especially in February, and in mid-June and July. The large values from February and July can be associated with arid dust intrusions from the Sahara desert, although they are more common during spring and sum-

mer.

Assessing the current legislation (*EC* 2008), we have analyzed regions surpassing 25 µg m^{-3} PM2.5 (Fig. 10). None of these regions exceed the mean annual limit legislated. However, areas closest to the annual limit value are AMB and Valls, with annual means over 15 µg m^{-3}. The southeastern part of Catalonia shows annual means over 10 µg m^{-3}, whereas the northern and some southwestern regions exhibit even lower concentrations.

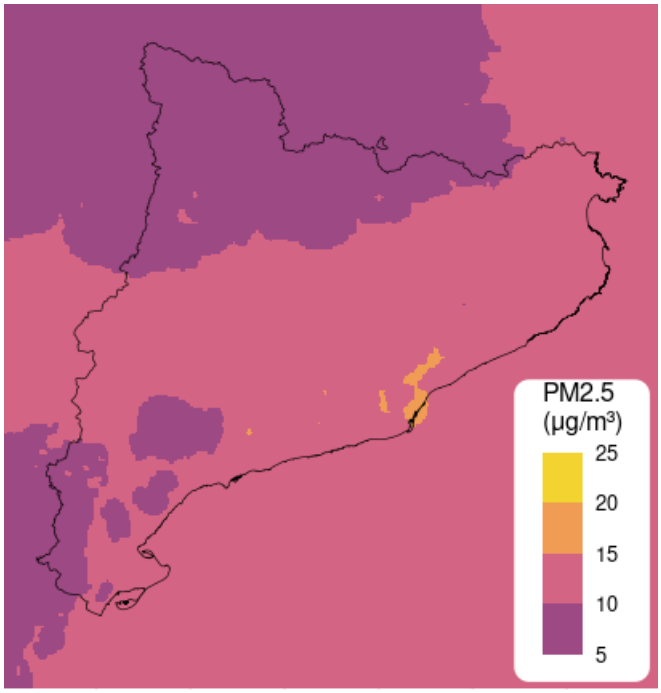

FIG. 10. Discrete scale for the annual PM2.5 mean concentration (µg m^{-3}) results obtained with data fusion (UK+GBM).

IV. CONCLUSIONS

We identified the Gradient Boosting Machine as the most effective method among those studied for filling in missing data in PM2.5 monitoring stations. We subsequently applied data fusion using both, filled and not filled datasets to evaluate its impact across various stations and the entire domain region.

Regarding the performance of gap-filling techniques, all techniques tended to overestimate PM2.5 values, except for GBM at Eixample's station, where it underestimated the observed values. GBM consistently showed a higher correlation coefficient comparing other techniques at both stations. However, in La Sènia's station, the combination of low data availability and isolated conditions negatively impacted GBM's predictive accuracy. Linear interpolation and persistence techniques exhibited a noticeable decrease in correlation coefficient as gap size increased, whereas GBM maintained relatively stable performance across different gap sizes.

We assessed the data fusion performance with and without GBM by conducting Leave-One-Out Cross-Validation at each station. Gap-filled observations improved data-fusion performance, particularly around the main Barcelona region when observation data availability is lower. Interpolated values from nearby stations and seasonal variations (Julian day) were key predictors influencing GBM's performance. Comparing concentrations post-data fusion with the raw model, we observed improvements in correlation coefficients, especially when GBM-filled data were used (+61%). This enhancement indicated better alignment with observed values and a significant reduction in mean bias and root mean squared error, even though GBM tended to underestimate values slightly.

No region exceeded the PM2.5 legal annual limit value set by Directive 2008/50/EC in 2019. However, AMB and Valls are closest to the threshold, with concentrations above 15 µg m^{-3}, and may reach future guidelines. The southeastern part of Catalonia recorded concentrations above 10 µg m^{-3}, while northern and some southwestern regions generally had lower levels.

To conclude, GBM has proven to be an effective technique for filling observational gaps. Compared to simpler techniques, GBM can provide useful gap-filled data even for extended data gaps. Furthermore, implementing data fusion with CALIOPE's model significantly enhances its performance. Correcting the model with observational data is relevant for accurately assessing air quality and improving our understanding of its distribution across Catalonia's region. In this manner, more accurate and localized governmental measures for reducing PM2.5 can be implemented.

ACKNOWLEDGMENTS

This work has received funding from the MePreCisa project of the UNICO I+D Cloud program that has the Ministry for Digital Transformation and of Civil Service and the EU-Next Generation EU as financing entities, within the framework of the PRTR and the MRR.

I would like to express my sincere gratitude to Cristina Carnerero and Jan Mateu for giving me the opportunity to work with them at the Barcelona Supercomputing Center. Their invaluable support and guidance throughout the project have been indispensable to its performance. I am also grateful to the Air Quality Services Team, whose ideas and support have greatly contributed to my professional development.

Moreover, I appreciate Mireia Udina for her constructive feedback and review of the study. Additionally, I wish to thank all the professors from the master for their teachings and attentive mentorship over the year.

Finally, I am thankful to my friends and family for their unwavering support and encouragement.

REFERENCES

Baldasano, J., M. Pay, O. Jorba, S. Gassó, and P. Jiménez-Guerrero, An annual assessment of air quality with the caliope modeling system over spain, *Science of The Total Environment*, *409*(11), 2163–2178, doi:https://doi.org/10.1016/j.scitotenv.2011.01.041, 2011.

Byun, D., and K. L. Schere, Review of the Governing Equations, Computational Algorithms, and Other Components of the Models-3 Community Multiscale Air Quality (CMAQ) Modeling System, *Applied Mechanics Reviews*, *59*(2), 51–77, doi:10.1115/1.2128636, 2006.

Caruana, R., and A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, in *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, p. 161–168, Association for Computing Machinery, New York, NY, USA, doi:10.1145/1143844.1143865, 2006.

Commission, E., Environment: Air, https://environment.ec.europa.eu/topics/air_en (last access: 19 June), 2024.

Cressie, N., *Statistics for Spatial Data*, chap. 1, pp. 1–26, John Wiley Sons, Ltd, doi:https://doi.org/10.1002/9781119115151.ch1, 1993.

EC, *Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe*, legislative Body: OP_DATPRO, 2008.

Friedman, J. H., Greedy function approximation: A gradient boosting machine., *The Annals of Statistics*, *29*(5), 1189 – 1232, doi:10.1214/aos/1013203451, 2001.

Greenwell, B. B. C. J., B., and G. Developers, gbm: Generalized boosted regression models, (https://CRAN. R-project.org/package=gbm, (last access: 15 May 2024), CRAN [code], R package version 2.1.8.1, 2022.

Guevara, M., C. Tena, M. Porquet, O. Jorba, and C. Pérez García-Pando, Hermesv3, a stand-alone multi-scale atmospheric emission modelling framework – part 1: global and regional module, *Geoscientific Model Development*, *12*(5), 1885–1907, doi:10.5194/gmd-12-1885-2019, 2019.

Hengl, H. G. B., T., and D. G. Rossiter, About regression-kriging: From equations to case studies, *Computational Geosciences*, pp. 1301–1315, doi:https://doi.org/10.1016/j.cageo.2007.05.001, 2007.

Horálek, D. B. e. a., J., Spatial mapping of air quality for european scale assessment, *Technology Representatives*, 2006.

Huang, Z. X. I. C. e. a., R., Air pollutant exposure field modeling using air quality model-data fusion methods and comparison with satellite aod-derived fields: application over north carolina, usa., *Air Qual Atmos Health*, *11*, 11–22, doi:10.1007/s11869-017-0511-y, 2018.

Institut_d'Estadística_de_Catalunya, gencat: Població a 1 de gener. províncies, https://www.idescat.cat/indicadors/?id=aecn=15223lang=es (last access: 19 June 2024), 2024.

Le Rest, K., D. Pinaud, P. Monestiez, J. Chadoeuf, and V. Bretagnolle, Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation, *Global Ecology and Biogeography*, *23*(7), 811–820, doi:https://doi.org/10.1111/geb.12161, 2014.

Lin, Y.-C., W.-J. Chi, and Y.-Q. Lin, The improvement of spatial-temporal resolution of pm2.5 estimation based on micro-air quality sensors by using data fusion technique, *Environment International*, *134*, 105,305, doi:https://doi.org/10.1016/j.envint.2019.105305, 2020.

Nickovic, S., G. Kallos, A. Papadopoulos, and O. Kakaliagou, A model for prediction of desert dust cycle in the atmosphere, *Journal of Geophysical Research: Atmospheres*, *106*(D16), 18,113–18,129, doi:https://doi.org/10.1029/2000JD900794, 2001.

Pardo-Iguzquiza, E., and M. Chica-Olmo, Geostatistics with the matern semivariogram model: A library of computer programs for inference, kriging and simulation, *Computers Geosciences*, *34*(9), 1073–1079, doi:https://doi.org/10.1016/j.cageo.2007.09.020, 2008.

Pérez, C., S. Nickovic, G. Pejanovic, J. M. Baldasano, and E. Özsoy, Interactive dust-radiation modeling: A step to improve weather forecasts, *Journal of Geophysical Research: Atmospheres*, *111*(D16), doi:https://doi.org/10.1029/2005JD006717, 2006.

Querol, X., A. Alastuey, S. Rodriguez, F. Plana, C. R. Ruiz, N. Cots, G. Massagué, and O. Puig, Pm10 and pm2.5 source apportionment in the barcelona metropolitan area, catalonia, spain, *Atmospheric Environment*, *35*(36), 6407–6419, doi:https://doi.org/10.1016/S1352-2310(01)00361-2, 2001.

Querol, X., et al., Lessons from the covid-19 air pollution decrease in spain: Now what?, *Science of The Total Environment*, *779*, 146,380, doi:https://doi.org/10.1016/j.scitotenv.2021.146380, 2021.

Skamarock, W. C., and J. B. Klemp, A time-split nonhydrostatic atmospheric model for weather research and forecasting applications, *Journal of Computational Physics*, *227*(7), 3465–3485, doi:https://doi.org/10.1016/j.jcp.2007.01.037, predicting weather, climate and extreme events, 2008.

Stein, C., Lectures on the theory of estimation of many parameters, *Journal of Soviet Mathematics*, *34*, 1373–1403, doi:https://doi.org/10.1007/BF01085007, 1986.

Su, Y., Prediction of air quality based on gradient boosting machine method, in *2020 International Conference on Big Data and Informatization Education (ICBDIE)*, pp. 395–397, doi:10.1109/ICBDIE50010.2020.00099, 2020.

WHO, Global air quality guidelines: particulate matter (pm2.5 and pm10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, world Health Organization, 2021.

Xing, Y.-F., Y.-H. Xu, M.-H. Shi, and Y.-X. Lian, The impact of pm2.5 on the human respiratory system, *Journal of Thoracic Disease*, *8*(1), 2016.

APPENDIX

TABLE III. Air Quality Monitoring Stations of XVPCA for PM2.5 observational data in 2019.

AQMS	Name station	Category	Availability (%)	Geographical coordinates
ES0392A	Manresa (CEIP La Font)	urban-suburban	47.1	(1.84 °E, 41.72 °N)
ES0559A	Barcelona (pl. Universitat)	urban-centre	90.7	(2.16 °E, 41.39 °N)
ES0567A	Barcelona (Zona Universitària)	urban-centre	93.7	(2.12 °E, 41.38 °N)
ES0691A	Barcelona (Poblenou)	urban-centre	92.3	(2.20 °E, 41.40 °N)
ES0692A	L'Hospitalet de Llobregat (av. Del Torrent Gornal)	urban-centre	49.3	(2.11 °E, 41.37 °N)
ES1123A	Constantí (Gaudí)	urban-suburban	47.1	(1.22 °E, 41.16 °N)
ES1148A	Sant Adrià del Besòs (Olímpic)	urban-centre	48.8	(2.22 °E, 41.43 °N)
ES1222A	Santa Maria de Palautordera (Martí Boada)	rural-regional	45.2	(2.44 °E, 41.69 °N)
ES1225A	Lleida (Irurita – Pius XII)	urban-centre	39.5	(0.62 °E, 41.62 °N)
ES1262A	Sabadell (Gran Via)	urban-centre	48.2	(2.10 °E, 41.56 °N)
ES1312A	Tarragona (Universitat Laboral)	urban-suburban	48.2	(1.21 °E, 41.10 °N)
ES1348A	Bellver de la Cerdanya (CEIP Mare de Déu de Talló)	rural-regional	46.0	(1.78 °E, 42.37 °N)
ES1438A	Barcelona (Eixample)	urban-centre	86.8	(2.15 °E, 41.39 °N)
ES1453A	Santa Coloma de Gramenet (Balldovina)	urban-centre	46.6	(2.21 °E, 41.45 °N)
ES1480A	Barcelona (Gràcia – Sant Gervarsi)	urban-centre	93.7	(2.15 °E, 41.40 °N)
ES1555A	Vilanova I la Geltrú (Ajuntament)	urban-centre	47.7	(1.73 °E, 41.22 °N)
ES1559A	La Bisbal d'Empordà	urban-suburban	41.1	(3.04 °E, 41.96 °N)
ES1642A	Vic (Estadi)	urban-suburban	43.6	(2.24 °E, 41.94 °N)
ES1663A	Sant Vicenç dels Horts (CEIP Mare de Déu del Rocío)	urban-suburban	44.4	(2.00 °E, 41.40 °N)
ES1684A	Rubí (ca n'Oriol)	urban-centre	48.8	(2.04 °E, 41.49 °N)
ES1754A	La Sènia	rural-near-city	45.2	(0.29 °E, 40.64 °N)
ES1841A	Mataró (Laboratori d'Aigües)	urban-suburban	47.1	(2.44 °E, 41.55 °N)
ES1851A	Berga (poliesportiu)	urban-suburban	46.3	(1.85 °E, 42.10 °N)
ES1852A	Barcelona (IES Goya)	urban-suburban	92.9	(2.17 °E, 41.42 °N)
ES1856A	Vandellòs I l'Hospitalet de l'Infant (viver)	urban-centre	92.1	(2.15 °E, 41.43 °N)
ES1891A	Granollers (Francesc Macià)	urban-centre	45.5	(2.29 °E, 41.60 °N)
ES1903A	Viladecans (Atrium)	urban-suburban	34.5	(2.01 °E, 41.31 °N)
ES1910A	Gavà (parc del Mil·leni)	urban-suburban	45.5	(1.99 °E, 41.30 °N)
ES1923A	Tona (zona esportiva)	rural-near-city	38.6	(2.22 °E, 41.85 °N)
ES1983A	El Prat de Llobregat (CEM Sagnier)	urban-suburban	44.1	(2.08 °E, 41.32 °N)