Science of the Total Environment 624 (2018) 189–200

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Air quality modelling in Catalonia from a combination of solar radiation, surface reflectance and elevation

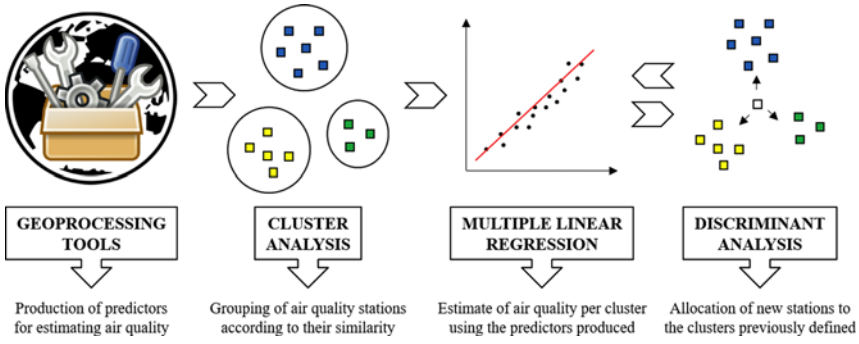
Daniel Jato-Espino [a,*], Elena Castillo-Lopez [b], Jorge Rodriguez-Hernandez [a], Francisco Ballester-Muñoz [a]

[a] GITECO Research Group, Universidad de Cantabria, Av. de los Castros 44, 39005 Santander, Spain
[b] Department of Geographical Engineering and Techniques of Graphic Expression, Universidad de Cantabria, Av. de los Castros 44, 39005 Santander, Spain

HIGHLIGHTS

- A new approach was conceived to model air quality in the region of Catalonia.
- The proposed methodology combined geoprocessing tools and multivariate statistics.
- Air quality was predicted from solar radiation, surface reflectance and elevation.
- The results provided highly accurate predictions of air quality at ungauged zones.
- The presence of irradiated built-up areas was found to endanger air quality.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 25 October 2017
Received in revised form 5 December 2017
Accepted 12 December 2017
Available online xxxx

Editor: P. Kassomenos

Keywords:
Air quality
Elevation
Geoprocessing tools
Multivariate statistics
Solar radiation
Surface reflectance

ABSTRACT

Air quality in developed areas is being increasingly compromised by the effect of urbanization, which is favouring the presence of atmospheric pollutants derived from human-induced activities. Land cover change is one of the consequences most closely associated with urbanization, leading to a growing presence of dark built-up surfaces. The target of this investigation was to model the Catalonian Air Quality Index (CAQI) from the combined effect of the surface reflectance capacity of urban surfaces with solar radiation and elevation. Geoprocessing tools were used to produce the information required to characterise these variables in the buffer areas surrounding 75 different air quality monitoring stations located across the region. Cluster analysis and Multiple Linear Regression (MLR) were applied to group these stations according to their similarity and replicate the annual mean values of CAQI recorded in Catalonia in 2011, respectively. Finally, discriminant analysis enabled assigning ungauged areas to the cluster and MLR model that best fitted their solar radiation, surface reflectance and elevation features. The implementation of this approach resulted in highly accurate predictions of CAQI, providing a mechanism of identification of areas having a number of days with poor air quality during the year. Since these areas were related to the presence of land cover types with high sunlight absorption, the proposed methodology was suggested to support the adoption of measures aimed at controlling urban air pollution based on replacing built-up surfaces by green infrastructure.

1. Introduction

Rapid population growth and urbanization are contributing to increasing air pollution in urban areas, particularly in developed countries (Han et al., 2016). Most of this pollution stems from human-related

* Corresponding author.
E-mail address: jatod@unican.es (D. Jato-Espino).

https://doi.org/10.1016/j.scitotenv.2017.12.139
0048-9697/

activities, such as energy consumption, industrialization or transportation, and is a source of risk to health, since it might eventually lead to cardiovascular and respiratory diseases (Andersen, 2017; Vardoulakis et al., 2003). In consequence, increasing attention is being paid to the impacts of urbanization on the environment, with emphasis on its alterations in terms of land cover (Alphan, 2003; Dewan and Yamaguchi, 2009) and solar radiation (Alpert and Kishcha, 2008; Wang et al., 2017) as potential threats for air pollution.

The development of land cover regression models for estimating air pollution has become a rich discipline in the field of atmospheric environment during the last 20 years. The pollutants addressed in these models included Nitrogen Oxides (NO_x) (Briggs et al., 2000; Gonzales et al., 2012; Muttoo et al., 2018; Stedman et al., 1997), Ozone (O_3) and Particulate Matter (PM) with diameter of <10 (PM_{10}) (Beelen et al., 2009) and 2.5 μm ($PM_{2.5}$) (Lee et al., 2017; Liu et al., 2016; Ross et al., 2007), Sulphur Dioxide (SO_2) (Amini et al., 2014) and Volatile Organic Compounds (VOC) (Wheeler et al., 2008). Besides a land cover-related variable, these studies considered other predictors, such as traffic, altitude, population, meteorology or precedent emissions. Their estimates reached coefficients of determination between 0.36 and 0.97, based on data recorded in a series of monitoring stations during sampling periods ranging from a few weeks to a whole year.

Specific investigations have also been conducted to explore the relationships between solar radiation and air pollutants. Gómez-Carracedo et al. (2015) suggested that the presence of photochemical reactions during the hours of maximum solar radiation favoured an increase in O_3, which coincided with the inferences achieved by Chou et al. (2007). In contrast, Shen et al. (2014) argued that O_3 was weakly correlated to solar irradiance due to the low concentrations of NO_x. Wang et al. (2005) found that secondary compounds of $PM_{2.5}$ exhibited high concentrations in summer as a result of strong solar radiation, a relationship which was confirmed by Vardoulakis and Kassomenos (2008) and Hajizadeh et al. (2017), who also reported positive correlations between solar radiation and NO_x, Carbon Monoxide (CO) and Benzene, Toluene, Ethylbenzene and Xylene (BTEX). However, these results differed from those yielded by a later study undertaken by Kassomenos et al. (2014), which indicated that the correlations between solar radiation and $PM_{2.5}$ and PM_{10} concentrations were not statistically significant.

The target of these works highlighted the lack of integrated approaches for evaluating air quality considering variables related to both land cover and solar radiation. To fill this gap, this research combined solar radiation, surface reflectance and elevation factors to model the Catalonian Air Quality Index (CAQI). This index was selected because its calculation includes the most commonly found air pollutants in urban daily life, such as CO, Nitrogen Dioxide (NO_2), O_3 and PM_{10} (Chen and Kan, 2008), and the data registered through its consideration is open access and widely available. In particular, the proposed methodology was tested and validated using the data recorded in 75 monitoring stations located in the region of Catalonia during 2011.

2. Methodology

The proposed methodology was conceived as a sequential combination of different multivariate statistical techniques, whose application was founded on data produced using Geographic Information Systems (GIS), as illustrated in Fig. 1. Hence, the set of factors or predictors identified as potential contributors to air pollution were generated through a series of geoprocessing tools framed within the discipline of spatial analysis. Then, a number of clusters were identified according to the values of these predictors across different air quality stations. This enabled maximising the prediction accuracy of the subsequent regression models built per cluster to estimate air quality. The last step concerned the application of discriminant analysis to validate the proposed approach by allocating different stations to the clusters whose regressions equations maximised their fit to observed values of air quality.

2.1. Framework

By 2017, Catalonia covered 32,106.5 km^2 and had 7477.131 inhabitants distributed among 947 municipalities (idescat.cat, 2017). This region is very dense and highly industrialised, circumstances that have favoured exceeding European air quality standards in 2015 and 2016 (Secció d'Immissions, 2015, 2016), especially in the Metropolitan Area of Barcelona, where two thirds of the population resided by 2017. The surface exposed to pollution levels above those legally permitted reached 24,000 km^2 in 2016, almost 75% of the whole area of Catalonia (Ceballos et al., 2015). These facts justified the need for developing new methods and approaches to help better manage air pollution in this region.

Air quality supervision in Catalonia is carried out by the Air Pollution Monitoring and Forecast Network, which consists of a series of stations aimed at measuring the levels of contamination reached across the region in relation to main atmospheric pollutants. Informing about the measurements of these stations per pollutant is a time-consuming and complex task, due to the technical details involved in the understanding and provision of these data. For this reason, a public information system based on an Air Quality Index (CAQI) was implemented in Catalonia since January 1995, in order to communicate population about the

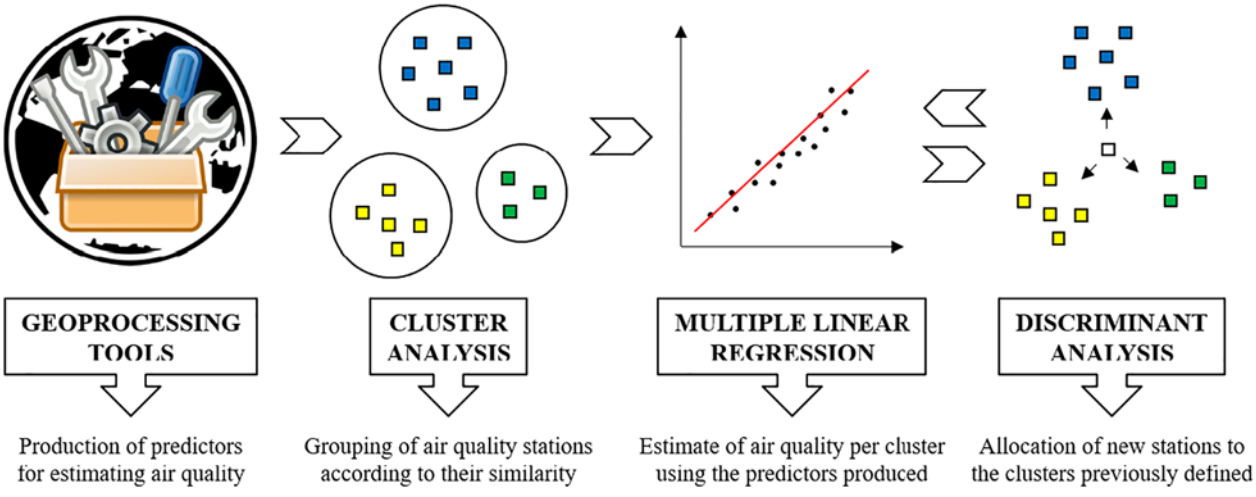


Fig. 1. Flowchart for the design and application of the proposed methodology to predict air quality.

quality of the air they breathe in a clear, concise and rapid manner (gencat.cat, 2015).

In addition to its ease of interpretation, the CAQI highlights by integrating the aspects considered in current European Union legislation. Furthermore, this approach provides an aggregated measure of the weighted contribution of the following pollutants to global air quality, which facilitates analysing its evolution with time: Ozone (O_3), Particulate Matter with diameter of <10 μm (PM_{10}), Carbon Monoxide (CO), Sulphur Dioxide (SO_2) and Nitrogen Dioxide (NO_2). The CAQI ranges from −100 to 100 and classifies air quality according to the three following levels: poor (<0), fair (0–49) and good (≥50) (gencat.cat, 2015). The approach taken to reach these thresholds consists of translating the concentration of the pollutants measured in the air quality monitoring stations located across Catalonia into a scale that indicates the effects of such contaminants on people, as specified in Table 1.

A good CAQI (≥50) has no negative effects on population health. In contrast, if the level of air quality decreases to fair (0–49), eye irritations and headaches are likely to occur. In these circumstances, the symptoms of heart and lung patients might be activated, whilst infants, the elderly and smokers may experience functional disorders in their respiratory and cardiovascular systems, such as increased respiratory rate, sensation of shortness of breath and palpitations. In those cases in which the CAQI is poor (<0), infants, the elderly and smokers could also suffer from inflammatory alterations in their respiratory system, including cough and bronchial spasms. In this situation, the generally healthy population can present functional disorders in their respiratory and cardiovascular systems as well, especially when practicing sports or other open-air physical activities.

The CAQI is not a mean obtained from the daily concentrations registered with respect to different pollutants, but a measure of the contaminant causing the greatest affection on air quality in a certain day (gencat.cat, 2015). Hence, the first step concerns the calculation of the value of CAQI associated with each pollutant recorded by the Catalonian Air Pollution Monitoring and Forecast Network, as represented in Table 1. Then, the value of CAQI for that day is determined as the lowest value of CAQI across all the pollutants.

The relationships between the values of CAQI recorded across Catalonia and solar radiation, surface reflectance and elevation were examined at the scale of circular buffer areas with a radius of 250 m around the air quality stations, a distance of influence which has been previously used to undertake spatial analyses at urban areas involving a monitoring network (van Hove et al., 2015). Therefore, the list of predictors required for estimating air quality (see Fig. 1) was produced according to the areas covered by such circular buffers.

Land cover maps were incorporated into the methodology to facilitate determining the values of surface reflectance associated with these buffers through the Albedo coefficient. This coefficient provides a measure of the amount of solar energy reflected from the Earth surface to space, being extremely linked to the role played by land cover to reduce temperatures near the ground (Bretz and Akbari, 1997; Taha, 1997). Hence, the identification of different land cover types was carried out with the support of the Spanish Land Use and Land Cover Information System (SIOSE, 2015). The SIOSE project is a land cover database reaching a level of detail four times higher than that of other European systems, such as the Corine Land Cover (CLC). Table 2 shows the 2-level classification of the SIOSE project, as well as the Albedo coefficients corresponding to each category, which were established based on the values found for urban covers in different studies (Coakley, 2003; Dobos, 2005; Wei et al., 2001).

The reference year for conducting the research was 2011, since the last available version of the SIOSE project was prepared by then. There were 75 valid air quality monitoring stations for that year distributed throughout Catalonia. Fig. 2 represents the location of the circular buffer areas associated with each of these stations, as well as a sample (Station ID: 74) of their land cover division according to the SIOSE classification. More than half of these stations were located in the south of the region, coinciding with the Metropolitan Area of Barcelona, where the highest levels of pollution in Catalonia are recorded due to increased urbanization and population (Ceballos et al., 2015). Still, there was a number of monitoring stations distributed across the rest of the region, including mountainous and rural areas, which enabled covering a variety of cases in terms of surface type configuration, solar radiation exposure and elevation.

2.2. Geoprocessing tools

Geoprocessing is any GIS-based operation related to the schemes and toolboxes that enable processing spatial data, such that inputs are

Table 1
Air quality classification according to the relationship between the Catalonian Air Quality Index (CAQI) and the pollutants measured in the monitoring stations located across the region.

Pollutant	Emission levels			
O_3; 1 h (μg/m^3)	0–110	111–180	181–240	>241
PM_{10}; 24 h (μg/m^3)	0–35	36–50	51–75	>76
CO; 8 h (mg/m^3)	0–5	6–10	11–15	>16
SO_2; 1 h (μg/m^3)	0–200	201–350	351–500	>501
NO_2; 1 h (μg/m^3)	0–90	91–200	201–400	>401
CAQI	100 to 50	49 to 0	−1 to −50	−51 to −100

Table 2
Land cover types and associated Albedo coefficients according to the Spanish Land Use and Land Cover Information System (Coakley, 2003; Dobos, 2005; SIOSE, 2015; Wei et al., 2001).

Group	SIOSE code	Land cover	Albedo
Artificial cover	EDF	Buildings	0.15
	ZAU	Artificial green areas and urban woodland	0.21
	LAA	Artificial water bodies	0.10
	VAP	Road, parking or pedestrian area without vegetation	0.08
	OCT	Other constructions	0.15
	SNE	Non-built soil	0.17
	ZEV	Extraction or discharge zones	0.17
Crops	CHA	Herbaceous: rice	0.18
	CHL	Herbaceous: other than rice	0.18
	LFC	Woody: citrus fruit	0.18
	LFN	Woody: non-citrus fruit	0.18
	LVI	Woody: vineyard	0.18
	LOL	Woody: olive grove	0.18
	LOC	Woody: other	0.18
	PRD	Meadows	0.19
Pastureland	PST	–	0.19
Woodland	FDC	Hardwood: deciduous	0.16
	FDP	Hardwood: perennials	0.16
	CNF	Coniferous	0.10
Brushwood	MTR	–	0.16
Barren vegetation	PDA	Beaches, dunes and sand	0.30
	SDN	Bare soil	0.17
	ZQM	Burned areas	0.17
	GNP	Glaciers and permanent snow	0.60
	RMB	Valleys	0.17
	ACM	Rocky: sea cliffs	0.17
	ARR	Rocky: rocky outcrops	0.17
	CCH	Rocky: screes	0.17
	CLC	Rocky: volcanic rocks	0.17
Humid cover	HPA	Inland wetlands: swamps	0.10
	HTU	Inland wetlands: peatbogs	0.10
	HAS	Inland wetlands: salt flats	0.10
	HMA	Marine wetlands: marshes	0.10
	HSM	Marine wetlands: salt marshes	0.10
Water cover	ACU	Inland waters: water courses	0.10
	ALG	Inland waters: lakes, ponds and reservoirs	0.10
	AEM	Inland waters: reservoirs	0.10
	ALC	Marine waters: coastal lagoons	0.10
	AES	Marine waters: estuaries	0.10
	AMO	Marine waters: seas and oceans	0.07

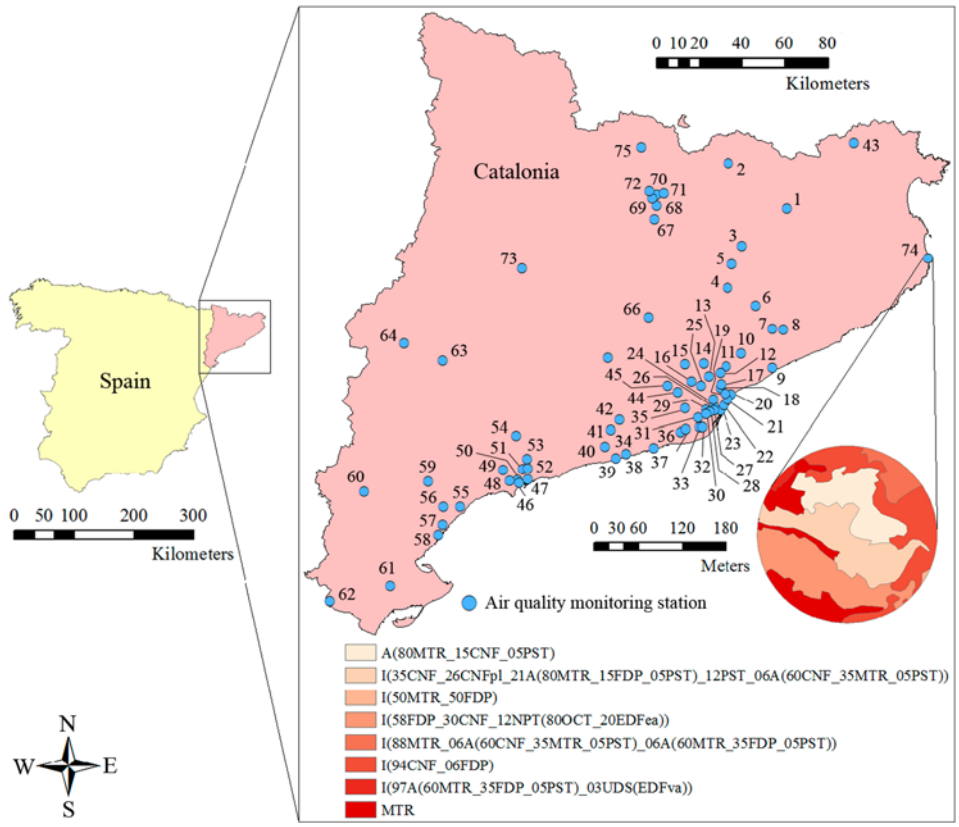


Fig. 2. Location of the 75 valid air quality stations in Catalonia in 2011 and zoom of the land cover classification of the circular buffer area associated with the monitoring station 74 according to the Spanish Land Use and Land Cover Information System (SIOSE).

manipulated to produce outputs containing relevant information. There is a vast array of different geoprocessing tools, among which some of the most widely used concern feature management, topology analysis and raster processing (ESRI, 2016a). In particular, the geoprocessing tools used to develop this research corresponded to those available in the desktop version of ArcGIS (ESRI, 2017), which provides a unified interface and structure for them all (Roberts et al., 2010).

The main data from which geographic tasks were carried out was the Digital Elevation Model (DEM) of Catalonia, which was acquired from the open data platform of the Spanish Geographic Institute (IGN in Spanish) in LiDAR (Light Detection and Ranging) format with a density of 0.5 points per m^2 (CNIG, 2017). The extraction of the LiDAR data, which was divided into files of 2 × 2 km extension, from the circular buffer areas depicted in Fig. 2 yielded the elevation information associated with the surroundings of each of the 75 monitoring stations considered.

In addition, the data included in the DEM provided the input required for applying the geoprocessing tools to conduct the solar radiation analysis of each buffer area. The procedure adopted in ArcGIS to carry out this analysis consists of calculating the insolation in a landscape from its variations in elevation, slope and shadows (ESRI, 2016b) during a certain time period, which in this case was the whole year 2011. Solar irradiance originated from the sun changes as it travels throughout the atmosphere, until being intercepted at the Earth surface in the form of direct, diffuse and reflected radiation (ESRI, 2016c).

Except for locations with an important presence of high reflective surfaces, reflected radiation is very small in comparison with direct and diffuse radiation. Therefore, ArcGIS omits this factor in the calculation of global radiation and applies the approach proposed by Fu and Rich (2003), based on calculating an upward-looking hemispherical view shed according to topographical data and overlapping it on direct sun and diffuse sky maps to determine direct and diffuse radiation, respectively (ESRI, 2016c). Fig. 3 illustrates this process for a single buffer area using the format of the ModelBuilder (ESRI, 2016d), the visual framework available in ArcGIS for creating geoprocessing workflows.

The application of the *Area Solar Radiation* tool yielded four outputs in raster format: "Global Radiation", "Diffuse Radiation", "Direct Radiation" and "Direct Duration". The input required by this tool was the DEM of the buffer area, whilst the parameters for establishing resolution, time configuration and hour interval were '200' (cells), 'Whole year with monthly interval' (2011) and '0.5' (hours), respectively. Then, the *Zonal Statistics* tool was used to calculate several descriptive statistics, including minimum (Min), maximum (Max), range, mean, standard deviation (SD) and sum. The inputs needed to run this operation were the "Radiation Outputs" determined before, which were compiled through the *Iterate Multivalue* tool, and the buffer area of the study station divided according to the SIOSE classification.

These zonal results enabled weighting the radiation outputs shown in Fig. 3 according to the Albedo coefficients associated with the land cover types contained in the buffer areas, using the values listed in Table 2. The addition of the DEM-based statistics for the buffer areas to the weighted radiation variables formed the set of predictors used to divide the 75 monitoring stations available in Catalonia in 2011 into clusters based on their similarity across these parameters.

2.3. Cluster analysis

The concept of cluster analysis was originally proposed by Tryon (1939) as a tool for grouping elements according to their similarity through the application of a series of theoretical methods. This technique is based on the working principle that the elements included in the same group are related to each other and unrelated to the elements allocated to other groups. Hence, cluster analysis was applied to this research to partition the list of air quality monitoring stations according to their similarity in terms of the solar radiation, surface reflectance and elevation conditions in their surroundings. This technique, as well as the

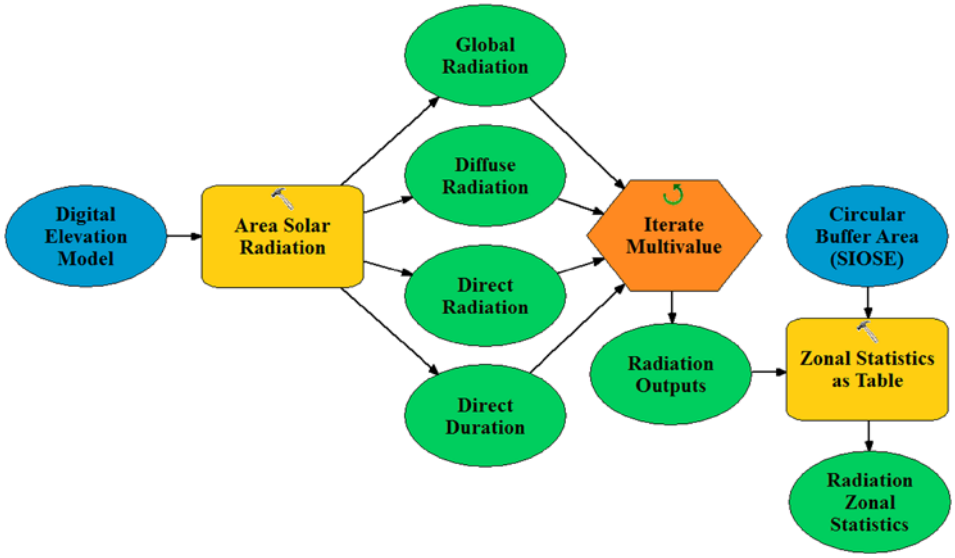


Fig. 3. Workflow for the production of predictors to estimate air quality.

other statistical methods included in the methodology, was applied with the support of the computer program Minitab (Minitab Inc., 2017).

The identification and composition of the number of cluster into which the whole set of stations should be divided was carried out through hierarchical clustering. This type of cluster analysis is based on calculating a distance matrix between the elements in the datasets, which enables identifying and grouping the two of them are closest to each other (Rokach and Maimon, 2005). The resulting cluster becomes indivisible, so that subsequent elements are grouped into increasingly large and heterogeneous conglomerates. Therefore, hierarchical clustering is an agglomerative procedure that ends with the creation of a single global cluster formed of all the elements contained in the dataset.

In particular, the number of clusters to consider was determined through the interpretation of dendrograms, which are tree-shape graphs representing the arrangement of conglomerates derived from hierarchical clustering. The results provided by this diagram depend on the linkage method used. Some approaches, including single, average and complete linkage, use any proximity measure, whilst some others require distances, such as the centroid, median and Ward's methods (Day and Edelsbrunner, 1984). Among them, the latter has been found to be very accurate when working with compact and spherical clusters (Blashfield, 1976; Hands and Everitt, 1987; Kuiper and Fisher, 1975), i.e. when the order of magnitude of the groups is similar.

Since this assumption was expected to be met due to the absence of outliers in the dataset, the Ward's method (Ward, 1963) was chosen for clustering. This approach seeks to minimise the variance within clusters by calculating the Sum of Squared Error (*SEE*) as formulated in Eq. (1), which measures the sum of the squared differences between each element and the centroid of the cluster to which they belong.

$$SEE = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \quad (1)$$

where n is the number of elements and x_i is the value of the ith element. Eq. (1) was applied iteratively, such that an element was selected to join a cluster at each stage if its inclusion minimised the sum of the intra-group distances. The clusters obtained as a result of this process enabled dividing the stations according to their similarity in relation to the outputs obtained from the application of the workflow depicted in Fig. 3, maximising the accuracy of the subsequent MLR models to be built for estimating values of CAQI.

2.4. Multiple Linear Regression

MLR was incorporated into the proposed methodology to determine the relationships between the mean values of CAQI recorded across the 75 air quality stations located in Catalonia in 2011 and the descriptive statistics related to solar radiation, surface reflectance and elevation obtained from geoprocessing. Eq. (2) provides the mathematical formulation to calculate the former as a linear combination of the latter:

$$y = b_0 + b_{1.1} * x_{1.1} + \ldots + b_{3.4} * x_{3.4} \ldots + b_{5.6} * x_{5.6} + e \quad (2)$$

where y is the response, $x_{i\ .\ j}$ stand for the value of the predictor i for the statistic j, b_0 and e represent the constant and the residuals, respectively, and $b_{i\ .\ j}$ indicate the relevance of the predictors in the MLR model. Table 3 shows the nomenclature used to characterise the list of predictors proposed in terms of the descriptive statistics calculated throughout the year of study as represented in Fig. 3.

The legitimacy of MLR was verified through the graphical inspection of the residuals contained in the models built, which were diagnosed in terms of linearity, independence, homoscedasticity and normality. Linearity was verified through the p-value of the regression term in the models, according to a significance level of 0.05 (Fisher, 1925). Plots of standardised residuals vs. station order and standardised predicted values enabled checking the assumptions of independence and homoscedasticity, respectively. Finally, a Quantile-Quantile (Q-Q) plot was used to guarantee the normality of residuals.

The validity of MLR analysis was double-checked to further ensure their validity to estimate future values of CAQI. On the one hand, the goodness-of-fit of the models built was assessed using the predicted R^2 coefficient, whose working principle is based upon systematically removing each air quality station from the model and then calculating how well the MLR equation predicts the omitted station. On the other hand, eight stations were excluded from the models at the beginning of this phase, in order to use them as a posteriori testers of the accuracy of discriminant analysis. Therefore, the MLR models produced at this stage played a primary role in the next step of the methodology, since their application enabled determining the reliability of the allocation process conducted to assign each excluded station to the cluster that best fitted their characteristics through discriminant analysis.

Table 3
List of predictors proposed for estimating the Catalonian Air Quality Index (CAQI).

Response	Predictor	Descriptive statistic					
		Min	Max	Range	Mean	SD	Sum
CAQI [−100, 100]	Weighted global radiation (Wh/m^2)	$x_{1.1}$	$x_{1.2}$	$x_{1.3}$	$x_{1.4}$	$x_{1.5}$	$x_{1.6}$
	Weighted diffuse radiation (Wh/m^2)	$x_{2.1}$	$x_{2.2}$	$x_{2.3}$	$x_{2.4}$	$x_{2.5}$	$x_{2.6}$
	Weighted direct radiation (Wh/m^2)	$x_{3.1}$	$x_{3.2}$	$x_{3.3}$	$x_{3.4}$	$x_{3.5}$	$x_{3.6}$
	Weighted direct duration (h)	$x_{4.1}$	$x_{4.2}$	$x_{4.3}$	$x_{4.4}$	$x_{4.5}$	$x_{4.6}$
	Elevation (m)	$x_{5.1}$	$x_{5.2}$	$x_{5.3}$	$x_{5.4}$	$x_{5.5}$	$x_{5.6}$

2.5. Discriminant analysis

Like clustering techniques, discriminant analysis belongs to the group of multivariate statistical methods devoted to the classification of elements into several groups. The main difference between discriminant and cluster analyses concerns the condition of the dataset used, since the former classifies elements from samples in which the groups are known aprioristically. Therefore, the purpose of discriminant analysis is to determine the membership of new elements based on the characteristics of the known groups (Lachenbruch, 1975).

Discriminant analysis can be seen as a logistic regression model in which a series of continuous independent variables are used to determine the membership group of the elements forming the dataset through according to a categorical dependent variable, such that the categories are equal to the groups (Press and Wilson, 1978). In analytical terms, the discriminant function D of two independent variables x_1 and x_2 is expressed as shown in Eq. (3), which enables differencing both groups as much as possible.

$$D = \beta_1 * x_1 + \beta_2 * x_2 \tag{3}$$

where β_1 and β_n are the weights of the independent leading the elements from both groups to achieve maximum and minimum scores in D, which results in the maximum possible separation between groups (SPSS, 2000). In other words, the application of this function must maximise and minimise the variance between and within groups, respectively. To this end, the values taken by D must be such that the distance d between the centroids $\overline{d}_1$ and $\overline{d}_2$ of both groups is maximised, according to Eqs. (4) and (5):

$$d = \overline{d}_1 - \overline{d}_2 \tag{4}$$

$$\overline{d}_i = \beta_1 * \overline{x}_1^{(i)} + \beta_2 * \overline{x}_2^{(i)} \tag{5}$$

where $\overline{d}_i$ is the centroid $\overline{d}$ of the group i, calculated by introducing the mean values of that group across the independent variables x_1 and x_2 in the discriminant function D (see Eq. (3)). To guarantee the validity of discriminant analysis, the groups must be differentiated in the independent variables beforehand, in order to enable finding a dimension in which the groups diverge from each other (SPSS, 2000). Otherwise, the centroids would be too close to each other, to the extent of making the distinction between the elements of both groups impossible.

The incorporation of discriminant analysis into the proposed methodology was aimed at serving as a means of allocation to enable determining which cluster must be assigned to new stations, in order to maximise the prediction accuracy of CAQI by applying the most convenient regression equation in each situation. Hence, the application of this technique to the case study of Catalonia in 2011 consisted of assigning the stations removed from the MLR step to the clusters that maximised their fit to the values of CAQI.

3. Results and discussion

This section compiles and discusses the results achieved from the application of the proposed methodology to the case study of Catalonia in 2011, following the four steps outlined in Fig. 1: Geoprocessing tools, Cluster analysis, Multiple Linear Regression (MLR) and Discriminant analysis. In addition, the section starts with an introduction focused on demonstrating the suitability of using the Catalonian Air Quality Index (CAQI) for developing this research.

3.1. Framework

The modelling and analysis of air quality in Catalonia was carried out on the basis of the annual mean values of CAQI obtained from the measures of the air pollutants included in Table 1 in 75 different monitoring stations located across the region in 2011. To prove the rigor and usefulness of using annual mean values of a measure like CAQI, the correlation coefficients between such values and the number of days within 2011 in which this parameter was good (>50), fair (0–49) and poor (<0) (see Table 1) were calculated. In particular, the Spearman's rho was the correlation coefficient selected for this purpose, since the aim was to find monotonic relationships between the values of CAQI and the number of days corresponding to each level.

The results collected in Table 4 demonstrated the statistical significance of these correlations (p-values < 0.05 in all cases), which were especially strong for the levels good and fair. The fact that several of the 75 stations considered recorded 0 days with poor CAQI explained the lower Spearman's rho reached for this level. Still, these results indicated that those stations with annual mean values in the range between 42 and 55 were poor in terms of CAQI several days throughout the year. Annual mean values ranging from 55 to 70 were very likely to be related to a number of days with fair CAQI levels, whilst those stations with values above 70 had generally good air quality all year round. Overall, these findings guaranteed the relevance of using this parameter as a measure of air quality, according to the emissions of O_3, PM_{10}, CO, SO_2 and NO_2 related to the three levels of CAQI (see Table 1). Furthermore, similar principles than those used for creating the CAQI have been implemented to create the World Air Quality Index (aqicn.org., 2017), an initiative providing air quality information in 600 major cities worldwide, based on data from >9000 stations. This project provides evidence of the potential applicability of the proposed approach, which might be easily extrapolated to hundreds of cities around the globe.

3.2. Geoprocessing tools

The application of the set of geoprocessing tools described in Fig. 3 to the buffer area zoomed in Fig. 2 (Station ID: 74) yielded the circular maps depicted in Fig. 4. The Digital Elevation Model (DEM) corresponding to this buffer area enabled running the solar radiation tool, which produced the outputs required for generating the predictors to be used in subsequent steps. Direct radiation was responsible for almost 80% of the global radiation in this particular buffer area, which explains the resemblance between both maps.

The circles shown in Fig. 4 provided information about the global solar radiation accumulated by each cell contained in the buffer area throughout 2011, broken down into diffuse radiation, direct radiation and direct duration. The processing of these values and those of elevation using zonal statistics, with the additional consideration of the zones established by the Spanish Land Use and Land Cover Information

Table 4
Spearman's correlation coefficients (rho) between the annual mean values of Catalonian Air Quality Index (CAQI) recorded in 2011 and the number of days during the year in which the air quality level was good, fair and poor.

Term 1	Term 2	Spearman's rho	p-Value
Annual mean value of CAQI [42, 96]	No. of days with good CAQI (≥50)	0.823	0.000
	No. of days with fair CAQI (0–49)	−0.832	0.000
	No. of days with poor CAQI (<0)	−0.505	0.000

System (SIOSE) in the case of radiation, enabled determining the minimum, maximum, range, mean, standard deviation and sum of these variables for each land cover type (see Table 2).

The last operation to end characterising the list of predictors proposed in Table 3 consisted of multiplying the radiation-related values per land cover type by their corresponding Albedo coefficients included in Table 2. This resulted in weighted values of radiation and duration, providing a measure of the weighted effect of radiation on air quality once the degree of surface reflectance of the ground surrounding the monitoring stations was taken into account.

3.3. Cluster analysis

The descriptive statistics obtained through the application of geoprocessing tools for the 75 buffer areas considered were used as the variables defining the clusters for grouping the monitoring stations according to their similarity. Since cluster analysis algorithms are based on calculating distances, these variables were previously standardised to avoid achieving misleading results due to the different scales in which they were measured.

Hence, the standardised variables were inputs to draw the dendrogram represented in Fig. 4 using the Ward method as linkage method and Euclidean metric as distance measure. The visual inspection of the dendrogram suggested that four clusters (CL1, CL2, CL3 and CL4) might be a suitable division (see Fig. 5). Considering a higher number of clusters would lead to unbalanced groups, whilst three or less cluster would result in a loss of accuracy. CL1 included stations located around the most urbanised areas in Catalonia and highlighted by reaching the highest values in direct duration. In contrast, CL2 and CL4 contained most of the stations located in Barcelona and its surroundings, resulting in low values of Albedo and elevation. Finally, CL3 consisted of the stations located in the north of the region, coinciding with mountainous areas and, therefore, high values of elevation.

Since cluster analysis was a preliminary step to develop regression equations for predicting CAQI, the number and composition of clusters was intended to maximise the goodness-of-fit of the MLR models to be built in the next step. Although the number of clusters identified was optimal, some stations included in each of these four groups were found to be improvable in terms of the goodness-of-fit of MLR, as a result of the differences between the working principles behind both statistical techniques.

3.4. Multiple Linear Regression

To palliate the lack of fit derived from cluster analysis, a cross-validation process was conducted to optimise the allocation of stations in the clusters identified in terms of MLR, such that each station was removed from its original cluster and added to the remaining clusters, in order to test how its omission or inclusion affected the prediction reliability of the MLR equations associated with every cluster. In the end, this task led to maximise the fit between measured and estimated

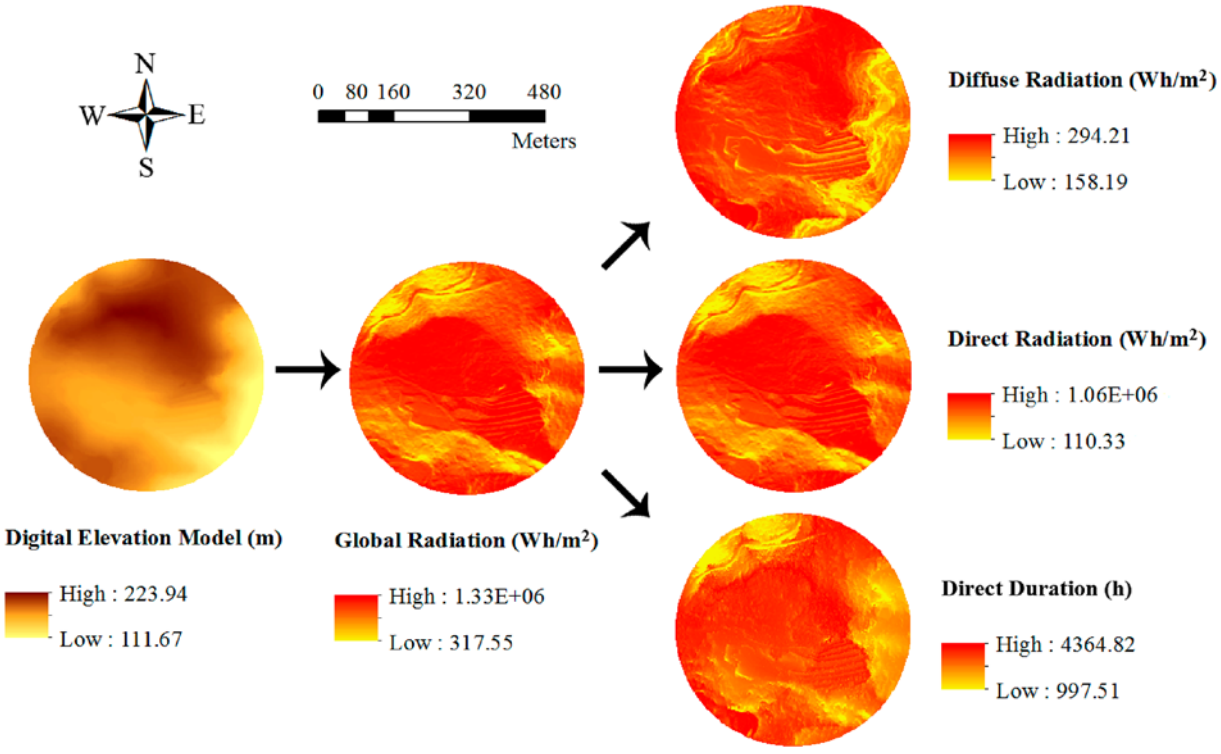


Fig. 4. Elevation (m), global radiation (Wh/m^2), diffuse radiation (Wh/m^2) and direct radiation (Wh/m^2) and duration (h) maps obtained for the circular buffer area corresponding to the monitoring station 74.

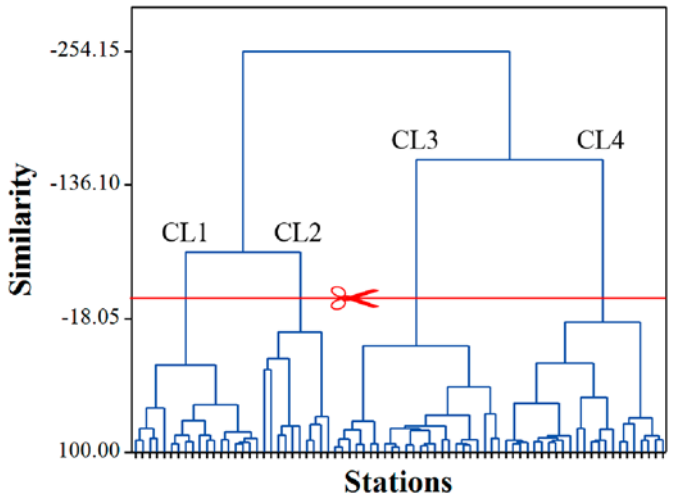


Fig. 5. Dendrogram for dividing the set of 75 monitoring stations into clusters according to their similarity in terms of solar radiation, surface reflectance and elevation.

CAQI, resulting in the following cluster arrangement: CL1–14 stations; CL2–19 stations; CL3–17 stations; CL4–25 stations.

Up to 8 stations were excluded from this step in proportion to the size of the clusters, in order to use them for checking the accuracy of discriminant analysis in the application of the last phase of the proposed methodology. Hence, 1, 2, 2 and 3 stations were omitted in the calculation of the MLR models corresponding to CL1, CL2, CL3 and CL4, respectively. Table 5 summarizes the main characteristics of these models, including the terms involved in their equations and the values achieved with respect to different goodness-of-fit measures.

The logic of the impacts of the Albedo coefficient, solar radiation and elevation on air quality, which have been documented in different studies over the years (Bisht et al., 2016; Peterson and Flowers, 1977; Touchaei et al., 2016; Twomey, 1974; U.S. EPA, 1978; Yamashita, 1973), were confirmed by Fig. 6. Dark surfaces involve lower values of Albedo and high sunlight absorption, increasing local temperatures and speeding up chemical reactions that diminish air quality, as depicted in Fig. 6a. Fig. 6b demonstrates that the relationship between weighted global radiation and CAQI is inversely proportional, which reaffirms the aforementioned positive impact of irradiated surfaces on air pollution. Moreover, since temperature decreases with elevation as molecules expand, the warming effect of solar radiation on air quality at high altitudes is mitigated as proved in Fig. 6c. According to these inferences, urban planning and management strategies concerning air quality should consider taking measures to replace part of the built-up skin of urban areas by lighter surfaces, preferably based on green infrastructure, in order to control atmospheric pollution and contribute to improving sustainability (Jato-Espino et al., 2017).

The standard R^2 coefficients shown in Table 5 suggested that >85% of the variations in the values of CAQI was explained by the solar radiation, surface reflectance and elevation-related predictors. Fig. 7 illustrates the excellent fit provided by the MLR models built, which demonstrated to be capable of predicting peaks and sinks of CAQI with high precision. The standard error of the regression S confirmed the accuracy of the estimates to replicate the values of CAQI measured in 2011. Furthermore, the adjusted and predicted R^2 coefficients reached guaranteed that the MLR models were not overfitted due to an excess of predictors and validated their use for making new predictions, respectively.

The residuals of these MLR models were analysed to fully ensure their reliability. The p-values of the regression term in Table 5, which were below the significance level in all cases (p-values < 0.05), guaranteed the linearity of residuals. Fig. 8 depicts the Q-Q plots, standardised residual versus fits plots and standardised residual versus order plots corresponding to the MLR models built for each cluster, which were used to verify the normality, homoscedasticity and independence of residuals, respectively. Since the points in the Q-Q plot lied close to a straight line, normality was assumed to be true. Moreover, the absence of marked trends and correlations in the standardised residual versus fits and order plots enabled accepting the hypotheses of homoscedasticity and independence, respectively. The adequate arrangement of the residuals in the plots included in Fig. 8 prevents the predictions and scientific insights achieved through the application of the MLR models represented in Table 5 from being biased or misleading. Consequently, these models were scientifically validated to estimate the CAQI in any location of the region.

3.5. Discriminant analysis

Since the regression equations built in the previous step were arranged according to clusters, the validation of the proposed methodology required a mechanism to allocate new study areas where to estimate air quality to one group or another, in order to ensure each location was assigned to the MLR model in Table 5 that maximised the fit of CAQI. This was accomplished through discriminant analysis, which was computed from group sizes, such that the likelihood of membership to a group increased as the size of the group increased. Furthermore, the stations considered for MLR were classified using a separate-groups

Table 5
Scheme of the Multiple Linear Regression (MLR) models built for estimating the Catalonian Air Quality Index (CAQI).

Term	CL1 (N = 13)		CL2 (N = 17)		CL3 (N = 15)		CL4 (N = 22)	
	Coef	p-Value	Coef	p-Value	Coef	p-Value	Coef	p-Value
Regression	–	0.001	–	0.000	–	0.000	–	0.000
b_0	318.35	0.000	71.89	0.000	136.61	0.000	149.90	0.000
$b_{1.3}$	1.01E−04	0.000	–	–	–	–	–	–
$b_{2.1}$	−1.18E−04	0.002	–	–	–	–	–	–
$b_{2.5}$	–	–	–	–	–	–	1.30E−03	0.000
$b_{3.1}$	–	–	−1.36E−05	0.001	−3.60E−04	0.000	–	–
$b_{3.4}$	−3.59E−05	0.033	–	–	–	–	−5.28E−05	0.000
$b_{4.1}$	–	–	–	–	–	–	4.92E−03	0.000
$b_{4.3}$	−3.43E−02	0.000	–	–	–	–	–	–
$b_{4.6}$	−5.43E−08	0.032	–	–	–	–	–	–
$b_{5.1}$	–	–	4.21E−02	0.000	−5.23E−02	0.000	−3.64E−02	0.006
$b_{5.2}$	−6.64E−02	0.001	–	–	–	–	–	–
$b_{5.5}$	–	–	3.85E−01	0.003	–	–	–	–
S		3.64		5.32		3.37		4.57
R^2		0.95		0.89		0.90		0.86
Adj. R^2		0.91		0.86		0.88		0.83
Pred. R^2		0.78		0.77		0.84		0.75

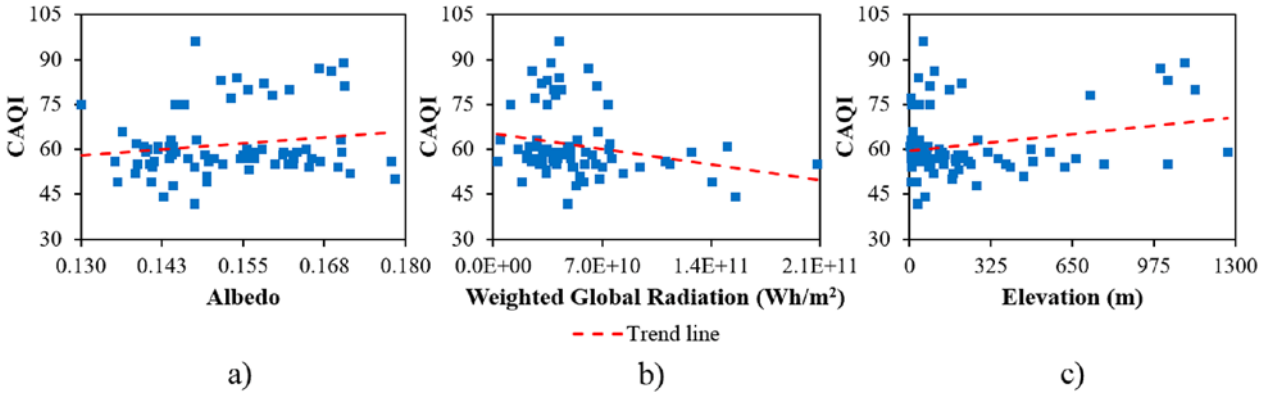


Fig. 6. Relationships between Air Quality Index (CAQI) and representative parameters of the proposed predictors: a) Albedo coefficient b) weighted global radiation (Wh/m^2) c) elevation (m).

covariance matrix based on discriminant functions, which resulted in the distribution depicted in Fig. 9.

Overall, 51 out of the 67 stations included in the MLR models schematised in Table 5 were allocated to their membership groups (74.63%), which involved that 3, 7, 2 and 5 stations that were supposed to be allocated to CL1, CL2, CL3 and CL4, respectively, were misassigned to any of the other three groups. The number of stations misallocated to their membership cluster was particularly high for CL2, due to the proximity of its centroid to those of CL3 and CL4, as demonstrated in Fig. 9. Regarding the stations removed from the MLR models, the allocation process undertaken using discriminant analysis and the subsequent application of the corresponding regression equations (see Table 5) yielded the results compiled in Table 6.

All the validation stations were assigned to the cluster that maximised their fit to the measured value of CAQI, except station 9, which was misassigned to CL4. Recalculations made using the equation corresponding to CL2, which was the second nearest group to station 9, yielded a predicted CAQI of 49.59, a value that resulted in an error in the order of magnitude of *S* for that MLR model (see Table 5). Consequently, discriminant analysis was demonstrated to provide very accurate results in the allocation process of new sites to the MLR models that maximised the prediction of CAQI according to the solar radiation, surface reflectance and elevation characteristics of their surrounding areas.

Overall, the results achieved proved the reliability and accuracy provided by the proposed methodology in the modelling of CAQI, such that its application might be used to predict air quality at non-monitored locations. Hence, the procedure to implement this methodology at ungauged sites would consist of (1) selecting a specific location, (2) delimiting a circular buffer area with a radius of 250 m around it, (3) computing the values of surface reflectance, solar radiation and elevation associated with such buffer area using geoprocessing tools, (4) assigning the location under study to the cluster that best fits its characteristics through discriminant analysis and (5) applying the MLR corresponding to such cluster to determine the value of CAQI being sought.

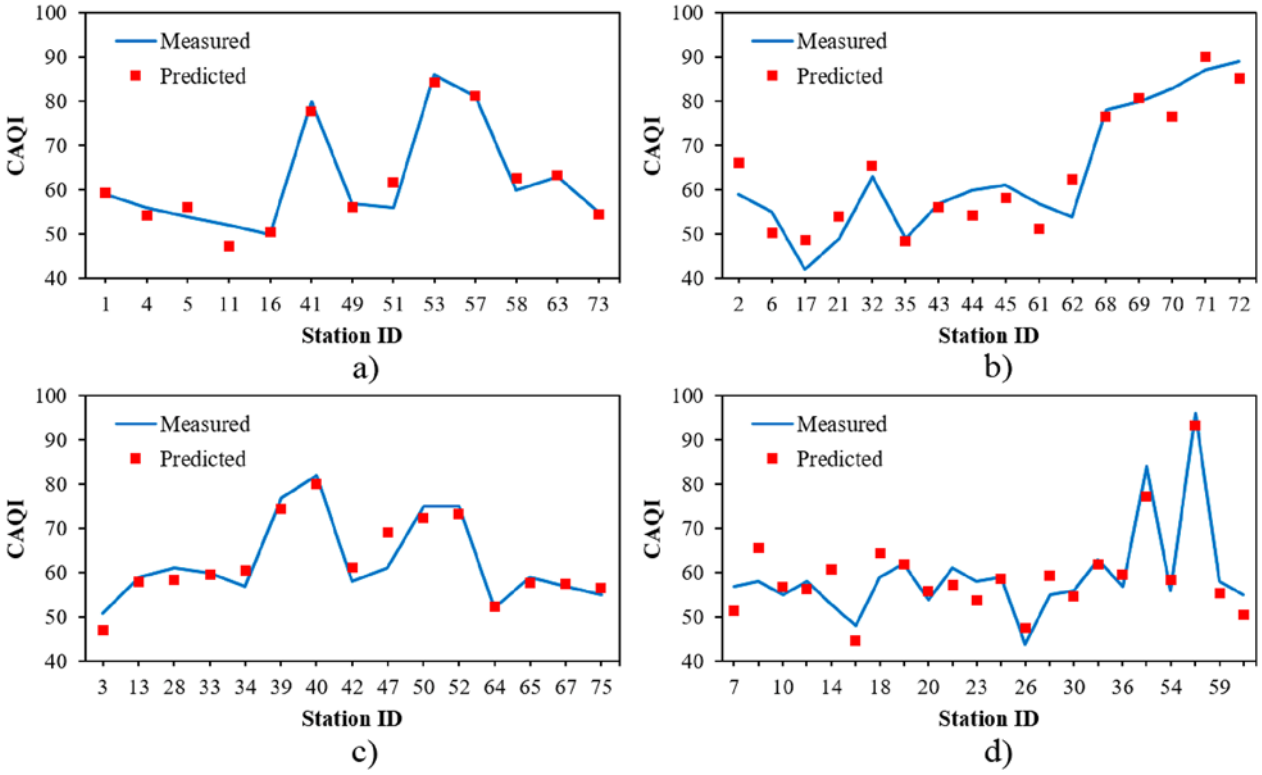


Fig. 7. Fit between measured and predicted values of Catalonian Air Quality Index (CAQI) in the stations included in a) Cluster 1 (CL1), b) Cluster 2 (CL2), c) Cluster 3 (CL3) and d) Cluster 4 (CL4).

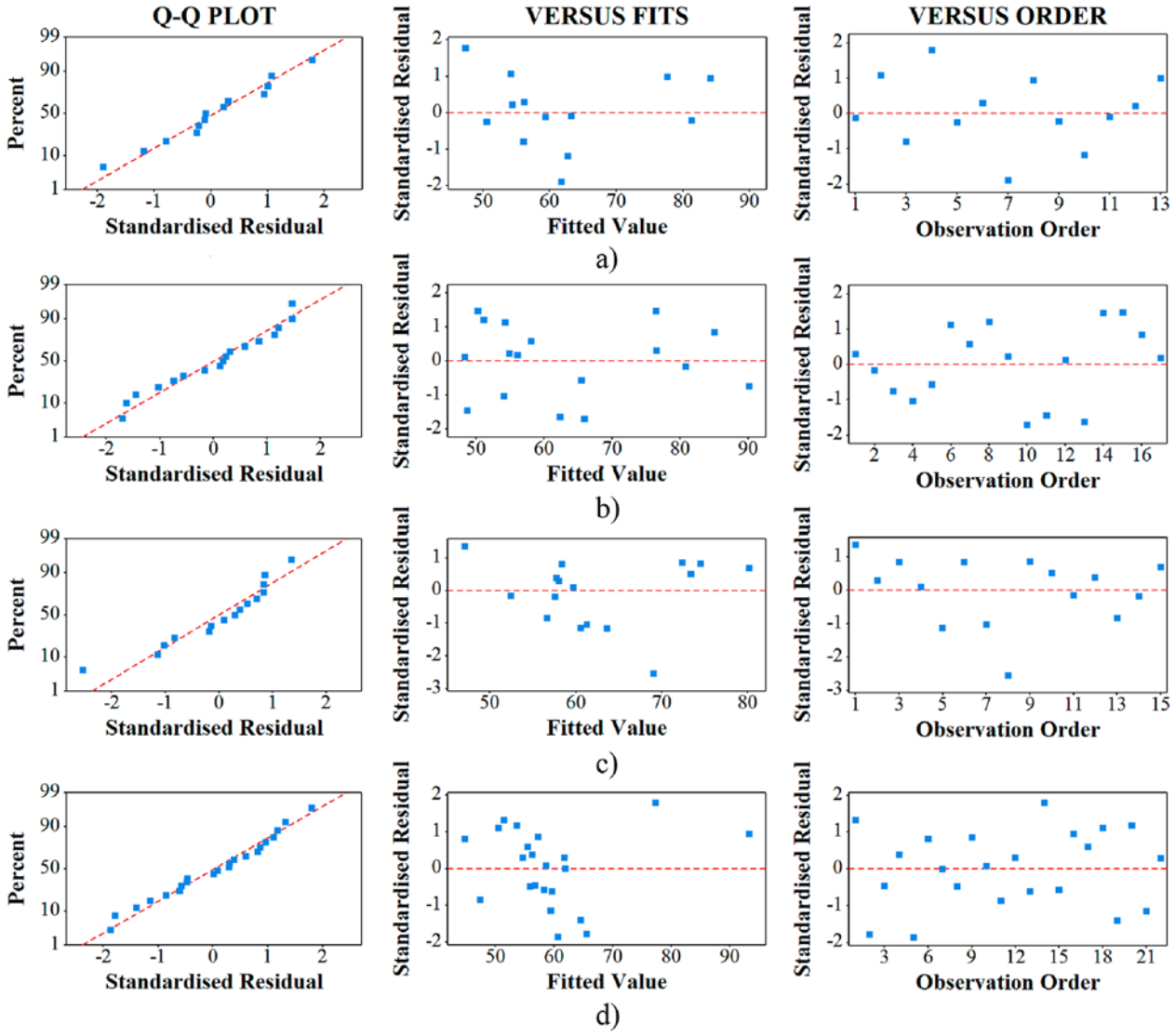


Fig. 8. Residual analysis for the Multiple Linear Regression (MLR) models to predict the Catalonian Air Quality Index (CAQI) in the stations included in a) Cluster 1 (CL1), b) Cluster 2 (CL2), c) Cluster 3 (CL3) and d) Cluster 4 (CL4).

4. Conclusions

This research demonstrated that the combination of geoprocessing tools and multivariate statistics can produce accurate estimates of air quality in Catalonia from the consideration of the interactions between the surface reflectance of urban surfaces with solar radiation and elevation. The results achieved through the sequential application of the different components included in the proposed methodology highlighted the synergetic role they played in the prediction of air quality.

Geoprocessing tools were found to provide a simple and accurate means to produce the solar radiation, surface reflectance and elevation predictors required to model air quality. Although cluster analysis did not provide an arrangement of the 75 monitoring stations located in Catalonia resulting in the maximisation of the fit between measured and predicted air quality, its combination with Multiple Linear

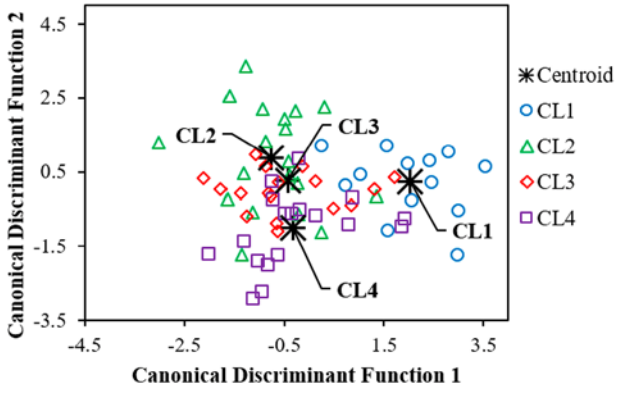


Fig. 9. Combined-groups plot representing the proximity of each monitoring station to the four clusters identified in previous steps.

Table 6
Comparison between measured and predicted values of Catalonian Air Quality Index (CAQI) for the validation stations using the Multiple Linear Regression (MLR) models corresponding to their assigned clusters through discriminant analysis.

Value	Station ID							
	9	25	27	38	46	48	56	60
Assigned cluster	4	3	4	4	3	4	2	1
Measured CAQI	54	60	49	66	75	59	60	57
Predicted CAQI	67.10	63.69	48.34	61.62	71.03	53.94	59.45	61.02

Regression (MLR) enabled replicating the values of Catalonian Air Quality Index (CAQI) recorded in 2011 with high precision. Finally, discriminant analysis facilitated the estimation of the levels of air quality at ungauged zones, since its application led to correctly assign more than three fourths of the stations analysed to its membership group, providing a reliable method to allocate new study areas to the cluster and corresponding MLR model best suited to their characteristics. In consequence, the stepwise implementation of the proposed methodology allowed estimating air quality with high precision, emerging as an effective tool to support decision-making processes in urban planning and management.

Since low annual mean values of CAQI were proved to be significantly correlated to the number of days within a year in which this variable is poor and, by extension, potentially harmful to human health, a tested procedure to estimate air quality with high accuracy like the proposed methodology can support the design of plan actions aimed at controlling air pollution in urban spaces. Furthermore, the negative effects of irradiated dark cover types on air quality found during the investigation suggested that substituting built-up surfaces by green infrastructure might reduce urban air pollution. Hence, future urban designs and restoration actions should be oriented to the implementation of technologies such as cool pavements and roofs and vegetated surfaces, in order to help mitigate the harmful impacts of urbanization on air quality through environmentally efficient land cover management practices.

Air quality indices are used as a recognised measure of atmospheric pollution in many cities worldwide, which guarantees the applicability of the methodology conceived in this research beyond the boundaries of Catalonia. However, although CAQI was statistically demonstrated to be a valid and rigorous indicator for air quality, future research in this line should also focus on implementing the proposed approach in the modelling of specific pollutants, such as Ozone (O_3), Particulate Matter (PM_{10}), Carbon Monoxide (CO), Sulphur Dioxide (SO_2) or Nitrogen Dioxide (NO_2).

Acknowledgments

This paper was possible thanks to the research project SUPRIS-SUReS (Ref. BIA2015-65240-C2-1-R MINECO/FEDER, UE), financed by the Spanish Ministry of Economy and Competitiveness with funds from the State General Budget (PGE) and the European Regional Development Fund (ERDF). The authors also wish to express their gratitude to all the entities that provided the data necessary to develop this study: the Department of Statistics and the Directorate of Environmental Monitoring of the Barcelona City Council, the Cartographic and Geological Institute of the Government of Catalonia and the National Centre of Geographic Information (CNIG) of the Spanish Ministry of Public Works and Transport.

References

gencat.cat, 2015. Qué es el Índice Catalán de Calidad del Aire (ICQA)? http://mediambient.gencat.cat/es/05_ambits_dactuacio/atmosfera/qualitat_de_laire/avaluacio/icqa/que_es_lindex_catala_de_qualitat_de_laire/ (2017)

idescat.cat, 2017. Anuario estadístico de Catalunya. Población a 1 de Enero. Provincias. https://www.idescat.cat/pub/?id=aec&n=245&lang=es2017.

aqicn.org., 2017. Worldwide Air Quality. http://aqicn.org/here2017.

Alpert, P., Kishcha, P., 2008. Quantification of the effect of urbanization on solar dimming. Geophys. Res. Lett. 35, L08801.

Alphan, H., 2003. Land-use change and urbanization of Adana, Turkey. Land Degrad. Dev. 14, 575–586.

Amini, H., Taghavi-Shahri, S.M., Henderson, S.B., Naddafi, K., Nabizadeh, R., Yunesian, M., 2014. Land use regression models to estimate the annual and seasonal spatial variability of sulfur dioxide and particulate matter in Tehran, Iran. Sci. Total Environ. 488–489, 343–353.

Andersen, M.S., 2017. Co-benefits of climate mitigation: counting statistical lives or life-years? Ecol. Indic. 79, 11–18.

Beelen, R., Hoek, G., Pebesma, E., Vienneau, D., de Hoogh, K., Briggs, D.J., 2009. Mapping of background air pollution at a fine spatial scale across the European Union. Sci. Total Environ. 407, 1852–1867.

Bisht, D.S., Tiwari, S., Dumka, U.C., Srivastava, A.K., Safai, P.D., Ghude, S.D., et al., 2016. Tethered balloon-born and ground-based measurements of black carbon and particulate profiles within the lower troposphere during the foggy period in Delhi, India. Sci. Total Environ. 573, 894–905.

Blashfield, R.K., 1976. Mixture model tests of cluster analysis: accuracy of four agglomerative hierarchical methods. Psychol. Bull. 83, 377–388.

Bretz, S.E., Akbari, H., 1997. Long-term performance of high-albedo roof coatings. Energ. Buildings 25 (2), 159–167.

Briggs, D.J., De Hoogh, C., Gulliver, J., Wills, J., Elliott, P., Kingham, S., et al., 2000. A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. Sci. Total Environ. 253, 151–167.

Ceballos, M.A., Segura, P., Gutiérrez, E., Diéguez, M., Barbé, S., Senán, J., et al., 2015. La calidad del aire en el Estado español durante 2015 Madrid (España): Ecologistas en Acción.

Chen, B., Kan, H., 2008. Air pollution and population health: a global challenge. Environ. Health Prev. Med. 13, 94–101.

Chou, C.C., Lee, C., Chen, W., Chang, S., Chen, T., Lin, C., et al., 2007. Lidar observations of the diurnal variations in the depth of urban mixing layer: a case study on the air quality deterioration in Taipei, Taiwan. Sci. Total Environ. 374, 156–166.

CNIG, 2017. Centro de descargas - Centro Nacional de Información Geográfica. http://centrodedescargascniges/CentroDescargas/locale?request_locale=en2017.

Coakley, J.A., 2003. Reflectance and albedo, surface. In: Holton, J.R., Curry, J.A., Pyle, J.A. (Eds.), Encyclopedia of Atmospheric Sciences. Academic Press, Cambridge, Massachusetts (U.S.), pp. 1914–1923.

Day, W.H.E., Edelsbrunner, H., 1984. Efficient algorithms for agglomerative hierarchical clustering methods. J. Classif. 1, 7–24.

Dewan, A.M., Yamaguchi, Y., 2009. Land use and land cover change in Greater Dhaka, Bangladesh: using remote sensing to promote sustainable urbanization. Appl. Geogr. 29, 390–401.

Dobos, E., 2005. Albedo. In: Lal, R. (Ed.), Encyclopedia of Soil Science. CRC Press, New York (U.S.), pp. 64–66.

ESRI, 2016a. Geoprocessing Tools. http://desktop.arcgis.com/en/arcmap/103/main/analyze/geoprocessing-tools.htm (2017).

ESRI, 2016b. Understanding Solar Radiation Analysis. http://desktop.arcgis.com/en/arcmap/103/tools/spatial-analyst-toolbox/understanding-solar-radiation-analysis.htm (2017).

ESRI, 2016c. Modeling Solar Radiation. http://desktop.arcgis.com/en/arcmap/103/tools/spatial-analyst-toolbox/modeling-solar-radiation.htm (2017).

ESRI, 2016d. What Is ModelBuilder? http://desktop.arcgis.com/en/arcmap/103/analyze/modelbuilder/what-is-modelbuilder.htm (2017)

ESRI, 2017. ArcGIS Desktop. http://desktop.arcgis.com/en/2017.

Fisher, R.A., 1925. Statistical Methods for Research Workers. Cosmo Publications, Edinburgh (Scotland).

Fu, P., Rich, P.M., 2003. A geometric solar radiation model with applications in agriculture and forestry. Comput. Electron. Agric. 37, 25–35.

Gómez-Carracedo, M.P., Andrade, J.M., Ballabio, D., Prada-Rodríguez, D., Muniategui-Lorenzo, S., Consonni, V., et al., 2015. Impact of medium-distance pollution sources in a Galician suburban site (NW Iberian Peninsula). Sci. Total Environ. 512–513, 114–124.

Gonzales, M., Myers, O., Smith, L., Olvera, H.A., Mukerjee, S., Li, W., et al., 2012. Evaluation of land use regression models for NO_2 in El Paso, Texas, USA. Sci. Total Environ. 432, 135–142.

Hajizadeh, Y., Mokhtari, M., Faraji, M., Mohammadi, A., Nemati, S., Ghanbari, R., et al., 2017. Trends of BTEX in the central urban area of Iran: a preliminary study of photochemical ozone pollution and health risk assessment. Atmos. Pollut. Res. (in press).

Han, L., Zhou, W., Pickett, S.T.A., Li, W., Li, L., 2016. An optimum city size? The scaling relationship for urban population and fine particulate (PM2.5) concentration. Environ. Pollut. 208, 96–101.

Hands, S., Everitt, B., 1987. A Monte Carlo study of the recovery of cluster structure in binary data by hierarchical clustering techniques. Multivar. Behav. Res. 22, 235–243.

van Hove, L.W.A., Jacobs, C.M.J., Heusinkveld, B.G., Elbers, J.A., Van Driel, B.L., Holtslag, A.A.M., 2015. Temporal and spatial variability of urban heat island and thermal comfort within the Rotterdam agglomeration. Build. Environ. 83, 91–103.

Jato-Espino, D., Yiwo, E., Rodriguez-Hernandez, J., Canteras-Jordana, J.C., 2017. Design and application of a Sustainable Urban Surface Rating System (SURSIST). Ecol. Indic. (under review).

Kassomenos, P.A., Vardoulakis, S., Chaloulakou, A., Paschalidou, A.K., Grivas, G., Borge, R., et al., 2014. Study of PM10 and PM2.5 levels in three European cities: analysis of intra and inter urban variations. Atmos. Environ. 87, 153–163.

Kuiper, F.K., Fisher, L., 1975. A Monte Carlo comparison of six clustering procedures. Biometrics 31, 777–783.

Lachenbruch, P.A., 1975. Discriminant Analysis. Hafner Press, New York (U.S.).

Lee, M., Brauer, M., Wong, P., Tang, R., Tsui, T.H., Choi, C., 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.

Liu, C., Henderson, B.H., Wang, D., Yang, X., Peng, Z., 2016. A land use regression application into assessing spatial variation of intra-urban fine particulate matter (PM2.5) and nitrogen dioxide (NO_2) concentrations in City of Shanghai, China. Sci. Total Environ. 565, 607–615.

Minitab Inc., 2017. Minitab® 17. http://www.minitab.com/en-us/products/minitab/2017.

Muttoo, S., Ramsay, L., Brunekreef, B., Beelen, R., Meliefste, K., Naidoo, R.N., 2018. Land use regression modelling estimating nitrogen oxides exposure in industrial south Durban, South Africa. Sci. Total Environ. 610–611, 1439–1447.

Peterson, J.T., Flowers, E.C., 1977. Interactions between air pollution and solar radiation. Sol. Energy 19, 23–32.

Press, S.J., Wilson, S., 1978. Choosing between logistic regression and discriminant analysis. J. Am. Stat. Assoc. 73, 699–705.

Roberts, J.J., Best, B.D., Dunn, D.C., Treml, E.A., Halpin, P.N., 2010. Marine geospatial ecology tools: an integrated framework for ecological geoprocessing with ArcGIS, Python, R, MATLAB, and C++. Environ. Model. Softw. 25, 1197–1207.

Rokach, L., Maimon, O., 2005. Clustering methods. In: Maimon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA (U.S.), pp. 321–352.
Ross, Z., Jerrett, M., Ito, K., Tempalski, B., Thurston, G.D., 2007. A land use regression for predicting fine particulate matter concentrations in the New York City region. Atmos. Environ. 41, 2255–2269.
Secció d'Immissions, 2015. La qualitat de l'aire a Catalunya. Anuari 2015 - Resum Barcelona (Spain): Generalitat de Catalunya. Direcció General de Qualitat Ambiental.
Secció d'Immissions, 2016. La qualitat de l'aire a Catalunya. Anuari 2016 - Resum Barcelona (Spain): Generalitat de Catalunya. Direcció General de Qualitat Ambiental.
Shen, Z., Cao, J., Zhang, L., Zhao, Z., Dong, J., Wang, L., et al., 2014. Characteristics of surface O_3 over Qinghai Lake area in Northeast Tibetan Plateau, China. Sci. Total Environ. 500–501, 295–301.
SIOSE, 2015. Sistema de Información de Ocupación del Suelo en España - Documento Técnico SIOSE 2011. D G Instituto Geográfico Nacional Servicio de Ocupación del Suelo S G de Cartografía. vol. 1.1 pp. 1–14.
SPSS, 2000. Guía para el análisis de datos el SPSS Madrid (Spain): Hispanoportuguesa SPSS.
Stedman, J.R., Vincent, K.J., Campbell, G.W., Goodwin, J.W.L., Downing, C.E.H., 1997. New high resolution maps of estimated background ambient NO(x) and NO_2 concentrations in the U.K. Atmos. Environ. 31, 3591–3602.
Taha, H., 1997. Urban climates and heat islands: albedo, evapotranspiration, and anthropogenic heat. Energ. Buildings 25, 99–103.
Touchaei, A.G., Akbari, H., Tessum, C.W., 2016. Effect of increasing urban albedo on meteorology and air quality of Montreal (Canada) - episodic simulation of heat wave in 2005. Atmos. Environ. 132, 188–206.
Tryon, R., 1939. Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers Malloy, Ann Arbor, Michigan (U.S.).
Twomey, S., 1974. Pollution and the planetary albedo. Atmos. Environ. (1967) 8, 1251–1256.
U.S. EPA, 1978. Altitude as a Factor in Air Pollution. U S Environmental Protection Agency, pp. 9–12 EPA-600/9-78-015.
Vardoulakis, S., Kassomenos, P., 2008. Sources and factors affecting PM10 levels in two European cities: implications for local air quality management. Atmos. Environ. 42, 3949–3963.
Vardoulakis, S., Fisher, B.E.A., Pericleous, K., Gonzalez-Flesca, N., 2003. Modelling air quality in street canyons: a review. Atmos. Environ. 37, 155–182.
Wang, Y., Zhuang, G., Tang, A., Yuan, H., Sun, Y., Chen, S., et al., 2005. The ion chemistry and the source of PM2.5 aerosol in Beijing. Atmos. Environ. 39, 3771–3784.
Wang, Y., Wild, M., Sanchez-Lorenzo, A., Manara, V., 2017. Urbanization effect on trends in sunshine duration in China. Ann. Geophys. 35, 839–851.
Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58, 236–244.
Wei, X., Hahmann, A.N., Dickinson, R.E., Yang, Z., Zeng, X., Schaudt, K.J., et al., 2001. Comparison of albedos computed by land surface models and evaluation against remotely sensed data. J. Geophys. Res.-Atmos. 106, 20687–20702.
Wheeler, A.J., Smith-Doiron, M., Xu, X., Gilbert, N.L., Brook, J.R., 2008. Intra-urban variability of air pollution in Windsor, Ontario-measurement and modeling for human exposure assessment. Environ. Res. 106, 7–16.
Yamashita, S., 1973. Air pollution study from measurements of solar radiation. Arch. Meteor. Geophy. B. 21, 243–253.