Heliyon 6 (2020) e04794

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Spatio-temporal air pollution modelling using a compositional approach

Joseph Sánchez-Balseca *, Agustí Pérez-Foguet

Research Group on Engineering Sciences and Global Development (EScGD), Civil and Environmental Engineering Department, Universitat Politècnica de Catalunya – BarcelonaTech (UPC), Spain

ARTICLE INFO

Keywords:
Statistics
Engineering
Atmospheric science
Environmental analysis
Environmental chemical engineering
Environmental impact assessment
Compositional data
CoDa
Air quality
Environmental statistics
Modelling

ABSTRACT

Air pollutant data are compositional in character because they describe quantitatively the parts of a whole (atmospheric composition). However, it is common to use air pollutant concentrations in statistical models without considering this characteristic of the data and, therefore, without control of common statistical problems, such as spurious correlations and subcompositional incoherence. This paper now proposes a daily multivariate spatio-temporal model with a compositional approach. The air pollution spatio-temporal model is based on a dynamic linear modelling framework with Bayesian inference. The novel modelling methodology was applied in an urban area for carbon monoxide (CO, $mg{\cdot}m^{-3}$), sulfur dioxide (SO_2, $\mu g{\cdot}m^{-3}$), ozone (O_3, $\mu g{\cdot}m^{-3}$), nitrogen dioxide (NO_2, $\mu g{\cdot}m^{-3}$), and particulate matter less than 2.5 μm in aerodynamic diameter ($PM_{2.5}$, $\mu g{\cdot}m^{-3}$). The proposal complemented and improved the conventional approach in air pollution modelling. The main improvements come from a fast multivariate data description, high spatial-correlation, and adequate modelling of air pollutants with high variability.

1. Introduction

Poor air quality in urban areas directly influences disease and decreases quality of life. Taking appropriate decisions in a timely period depends on the measurement and analysis of air parameters (Marinov et al., 2016). Air pollutants can be measured using either remote sense data (satellite images) or a monitoring network extended over a specific territory. The principal limitations of remote sense data are (a) the inverse relationship between spatial and temporal resolution, i.e., a satellite image could have high spatial resolution but low temporal resolution, and vice versa (Yao et al., 2018; Yang et al., 2019; Ban et al., 2020); (b) the satellite image often shows the vertical integration of air pollutants due to their passive remote sensing theory, causing issues in obtaining air pollutants near the surface (Zheng et al., 2017, 2018). The principal disadvantage of a monitoring network is the limited number of monitoring stations installed (low spatial resolution). Recent studies suggest the application of low-cost sensors in monitoring air pollutants, increasing the number of monitoring sites, and thus the spatial resolution; while the uncertainties in these observations remain an issue (Shi et al., 2018; Zhao et al., 2019). In both exposed cases, predictions from numerical models are used (Zannetti, 1990; Mayer, 1999; Dominici et al., 2002; Cetin et al., 2017; Paci, 2013; Vlachokostas et al., 2009; Huang et al., 2019). Further, these predictions are used in epidemiological researches (Mölter et al., 2010; Arroyo et al., 2019).

It is common to find spatio-temporal air quality models in the literature that use a univariate approach (Arakia et al., 2018; He et al., 2019; Hu et al., 2019; Pak et al., 2020). Dynamic linear models (DLM) are commonly used in air quality univariate modelling and they have been widely reviewed (Cocchi et al., 2007; Cameletti et al., 2011; Fassò and Finazzi, 2011; Sahu, 2012; Gutiérres et al., 2016; Shaddick et al., 2018). It is less common to find air quality models that use a multivariate approach (i.e., multiple response). This approach is frequently used in models related to air quality indices (AQI) (Jato-Espino et al., 2018; Zhang, 2019). Shaddick and Wakefield (2002) presented a multivariate daily spatio-temporal air pollution model for London using a dynamic linear modelling framework with Bayesian inference with the Markov chain Monte Carlo (MCMC) method. Blangiardo et al. (2019) presented a hierarchical model to assess multi-pollutant effects in time-series studies, using daily pollutants, weather, and mortality data.

Environmental sciences usually use closed data, i.e, they quantitatively represent the parts of a whole (e.g., μg/l, mg/kg, wt%) and only the proportions of their parts are assumed to be informative. If this information is not treated using a compositional approach, it could have severe consequences in the statistical analysis of the data (e.g., spurious correlations and subcompositional incoherence) (Filzmoser et al., 2010;

* Corresponding author.
E-mail address: joseph.sanchez@upc.edu (J. Sánchez-Balseca).

https://doi.org/10.1016/j.heliyon.2020.e04794
Received 24 April 2020; Received in revised form 26 June 2020; Accepted 24 August 2020

Table 1. Main parameters of monitoring stations.

Station Name	Location	Elevation (m.a.s.l.)	Station code
Carapungo	78°26′50″ W, 0°5′54″ S	2851	ST_1
Belisario	78°29′24″ W, 0°10′48″ S	2835	ST_2
El Camal	78°30′36″ W, 0°15′00″ S	2840	ST_3
Cotocollao	78°29′59,2″ W, 0°06′38,8″ S	2739	ST_4
Centro	78°30′50.4″ W, 0°13′17.6″ S	2453	ST_5

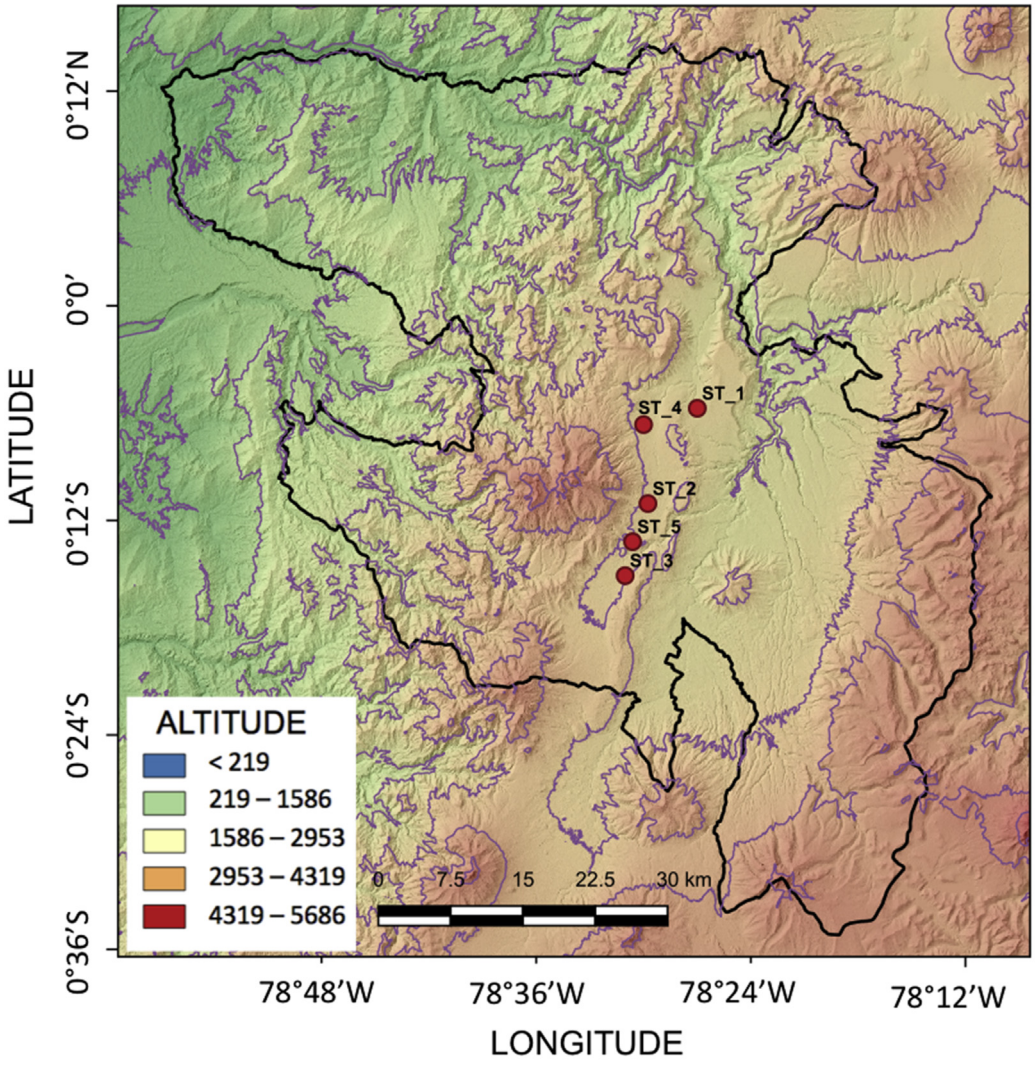

Figure 1. Orography of Quito. The five monitoring stations are indicated with red dots, and the administrative boundary of Quito is illustrated with a black polygon. The stations labels in the map refer to the "Station code" column in Table 1.

Egozcue et al., 2012). This is often neglected in statistical analysis, where a simple log transformation of the single variables is insufficient to put data into an acceptable geometry (Reimann et al., 2017). These problems in data with compositional character could be eliminated through log-ratio methods (Buccianti et al., 2006). For instance, Aitchison (1982) developed the additive–log-ratio (alr) and centred-log-ratio (clr) transformations; Egozcue et al. (2003) introduced the isometric-log-ratio (ilr) transformation.

Modelling and simulation works in air quality considering a compositional approach are still scarce (Sánchez-Balseca and Pérez-Foguet, 2019), with the few studies related to AQI (Jarauta-Bragulat et al., 2016; AL-Dhurafi et al., 2018). In contrast, this is not the case for issues related to water (Buccianti and Pawlowsky-Glahn, 2005; Blake et al., 2016; Owen et al., 2016; Pérez-Foguet et al., 2017; Ezbakhe and Pérez-Foguet, 2019; Quispe-Coica and Pérez-Foguet, 2020) or to soil (Reimann et al., 2012; Shi-wen et al., 2013; López-Abente et al., 2018; Petrika et al., 2018).

This article proposes a spatio-temporal daily multivariate air pollutant data using DLM, and Bayesian inference through MCMC methods with compositional data (CoDa) approach. The proposed model was applied using five monitoring stations with data of five air pollutants during six years in Quito, Ecuador. The remaining of this article provides the site description, datasets used, a brief background on statistical tools (DLM and compositional data analysis), and methodology (Section 2),

Table 2. Log-ratio coordinate definition.

Level	Log-ratio coordinates
1	$y_1^* = \sqrt{\frac{1\cdot 1}{1+1}}\ln\frac{x_1}{x_2}$
2	$y_2^* = \sqrt{\frac{1\cdot 2}{1+2}}\ln\frac{(x_1\cdot x_2)^{\frac{1}{2}}}{x_3}$
3	$y_3^* = \sqrt{\frac{1\cdot 3}{1+3}}\ln\frac{(x_1\cdot x_2\cdot x_3)^{\frac{1}{3}}}{x_4}$
4	$y_4^* = \sqrt{\frac{1\cdot 4}{1+4}}\ln\frac{(x_1\cdot x_2\cdot x_3\cdot x_4)^{\frac{1}{4}}}{x_5}$
5	$y_5^* = \sqrt{\frac{1\cdot 5}{1+5}}\ln\frac{(x_1\cdot x_2\cdot x_3\cdot x_4\cdot x_5)^{\frac{1}{5}}}{x_6}$

the results (Section 3), the discussion (Section 4), and the principle conclusions (Section 5).

2. Material and methods

2.1. Site description

Quito covers 4 230.6 km^2 and has a population of 2 505 344 inhabitants (EMASEO, 2011). The topography of the region is characterized by a complex terrain in a basin running in a NE to SW direction, surrounded by the Andes montains. It is situated in a narrow mountain valley at 2 800 m above sea level, and temperature inversions are common events (Jurado and Southgate, 1999).

2.2. Data

Datasets used were collected hourly from the monitoring network of Quito over six years (2009–2014). Samples with negative values of concentrations were discarded. To obtain the daily mean concentrations, at least 75% hourly data was required. The monitoring networking comprises five stations with distinct characteristics (Table 1, Figure 1). These stations measured with Thermo Fisher Scientific EPA standard methods both meteorological data (temperature and pressure) as well as five air pollutants: carbon monoxide (CO, mg·m^{-3}), sulphur dioxide (SO_2, μg·m^{-3}), ozone (O_3, μg·m^{-3}), nitrogen dioxide (NO_2, μg·m^{-3}), and particulate matter <2.5 μm ($PM_{2.5}$, μg·m^{-3}). The location and quality control processes of the monitoring stations were established by the Environmental Agency of Quito following the criteria for air quality monitoring as set by the Environmental Protection Agency of the United States (USEPA) (Secretaria de Ambiente del DMQ, 2017).

2.3. Background on statistical tools

2.3.1. Compositional data analysis

In mathematical terms, CoDa are represented as pertaining to a sample space called the simplex S^D (Equation 1):

$$S^D = \left\{ x = (x_1, x_2, x_D) : x_i > 0 (i = 1, 2, D), \sum_{i=1}^{D} X_i = K \right\}. \quad (1)$$

where K is a given positive constant, defined *a priori* and dependent on how the parts are measured (Buccianti, 2013). The elements of a composition, x_i, are called components or parts, and the only relevant information is contained in the ratios between components (Pawlowsky-Glahn et al., 2015). To obtain the log-ratio coordinates, the isometric log-ratio (ilr) transformation is applied (Equation 2):

$$\boldsymbol{y} = ilr(\boldsymbol{x}) = \ln(\boldsymbol{x}) \cdot \boldsymbol{V}. \quad (2)$$

where $\boldsymbol{x}$ is the vector with the D parts of the compositions, $\boldsymbol{V}$ a $D \times (D-1)$ matrix denotes the orthonormal basis in the simplex, and $\boldsymbol{y}$ is the vector with the $D-1$ log-ratio coordinates of the composition on the basis $\boldsymbol{V}$.

In this framework, the procedure of the sequential binary partition (SBP) to identify orthonormal coordinates was adopted (Egozcue and Pawlowsky-Glahn, 2005). A SBP denotes a hierarchy of the parts of a composition and contains successive splits of the parts into two groups, coded by the signs + and −. For instance, the components of $\boldsymbol{y}$ with respect to the basis $\boldsymbol{V}$ are:

$$y_i^* = \sqrt{\frac{r_i s_i}{r_i + s_i}} \ln\left(\frac{g_m(x_{i+})}{g_m(x_{i-})}\right); i = 1, \ldots, D-1 \quad (3)$$

where y_i^* is the ith orthonormal coordinate of the composition, $g_m(x_{i+})$ and $g_m(x_{i-})$ are the geometric mean of the components (coded as + and −, respectively) in the ith partition, and r_i and s_i are the number of components (coded as + and −, respectively) (Egozcue and Pawlowsky-Glahn, 2005). After the log-ratio coordinates are obtained, the conventional statistical tools can be applied. Finally, the simulate values (obtained from statistical tools) can be back-transformed to the original space using the inverse ilr operation, defined as:

$$\boldsymbol{x} = \mathscr{C}[\exp(\boldsymbol{V}\boldsymbol{y})]. \quad (4)$$

where $\boldsymbol{y}$ contains the log-ratio ilr coordinates of $\boldsymbol{x}$ with respect to the basis $\boldsymbol{V}$. The closure operator is defined as $\mathscr{C}$,

$$\mathscr{C}[\boldsymbol{x}] = \left(\frac{x_1}{\sum_{i=1}^{D} x_i}, \frac{x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{x_D}{\sum_{i=1}^{D} x_i}\right). \quad (5)$$

For a 6-part composition, $\boldsymbol{x} = (x_1, x_2, x_3, x_4, x_5, x_6)$, a SBP example can be:

Order	x_1	x_2	x_3	x_4	x_5	x_6	r	s
1	+1	−1	0	0	0	0	1	1
2	+1	+1	−1	0	0	0	2	1
3	+1	+1	+1	−1	0	0	3	1
4	+1	+1	+1	+1	−1	0	4	1
5	+1	+1	+1	+1	+1	−1	5	1

And thus, the orthonormal basis:

$$\boldsymbol{V} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{6} & 1/\sqrt{12} & 1/2\sqrt{5} & 1/\sqrt{30} \\ -1/\sqrt{2} & 1/\sqrt{6} & 1/\sqrt{12} & 1/2\sqrt{5} & 1/\sqrt{30} \\ 0 & -\sqrt{2/3} & 1/\sqrt{12} & 1/2\sqrt{5} & 1/\sqrt{30} \\ 0 & 0 & -\sqrt{3}/2 & 1/2\sqrt{5} & 1/\sqrt{30} \\ 0 & 0 & 0 & -2/\sqrt{5} & 1/\sqrt{30} \\ 0 & 0 & 0 & 0 & -\sqrt{5/6} \end{bmatrix}$$

The log-ratio ilr coordinates can be obtained using Eq. (3) (see Table 2). The log-ratio coordinates represent the influence of each part over the composition and could take into account the relationship between the components in order to obtain an optimal base (Meagher et al., 1967; Environmental Protection Agency, 2001).

2.3.2. Dynamic linear model

Let $\mathbf{y}_{spt}$ denote the observed pollutant concentration p at spatial location s (s = 1, …, S) on day t ($t = 1, \ldots, T$) and assume the observation equation as

$$\boldsymbol{y}_{spt} = \boldsymbol{X}_{spt} \cdot \boldsymbol{\beta} + \boldsymbol{\theta}_{spt} + \boldsymbol{V}_{spt}. \quad (6)$$

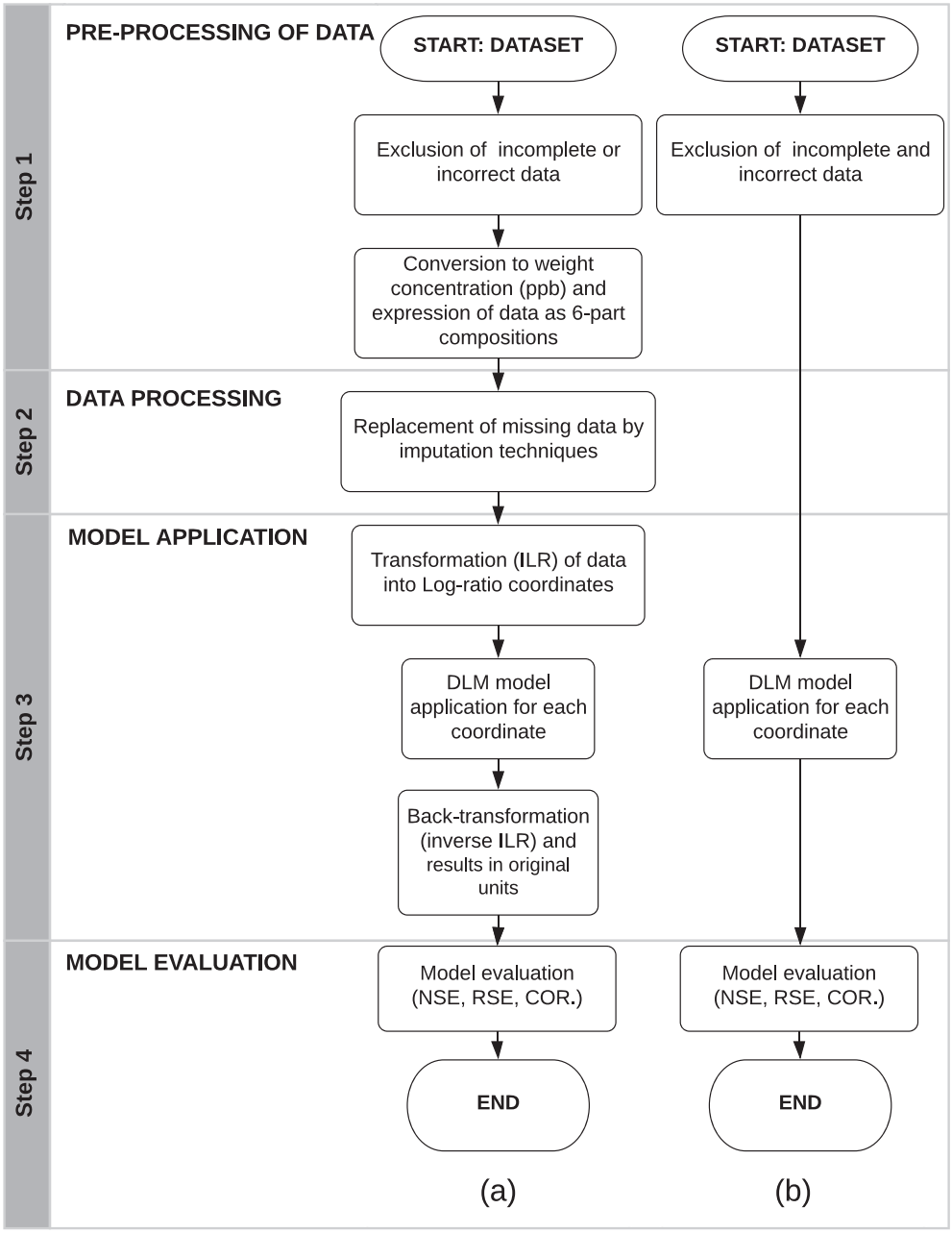

Figure 2. Algorithm of (a) the proposed approach, and (b) the conventional approach for spatio-temporal air pollution modelling using DLM.

where $\boldsymbol{V}_{spt}$ represents the measurement errors that are assumed to be independent and distributed $N(0, \sigma_v^2)$ (also named a Gaussian white-noise processes). The measurement error variance is also called the nugget effect (σ_v^2) (Cressie, 1993). The vector $\boldsymbol{\beta}$ is a vector of regression coefficients, and $\boldsymbol{X}_{spt}$ represents a vector of regressors that change temporally (large-scale component including meteorological and geographical covariates). The term $\boldsymbol{\theta}_{spt}$ is the realization of the latent spatio-temporal process (true unobserved levels of p pollutants on day t at site s) and is given by the dynamic autoregressive first-order model:

$$\boldsymbol{\theta}_{spt} = \boldsymbol{\theta}_{s,t-1} + \boldsymbol{w}_{pt} + \boldsymbol{m}_{ps}. \quad (7)$$

The last equation is termed the system equation. $\boldsymbol{w}_{pt}$ has a multivariate normal distribution, $MVN(0, \Sigma_p)$, with zero mean and variance–covariance matrix (Σ_p). This matrix contains variances σ_{wp}^2 and represents the covariance between pollutants (for more details, see Shaddick and Wakefield, 2002).

The site effects of pollutant p at station site s $\boldsymbol{m}_{ps}$ has also a multivariate normal distribution, $MVN(0, \Sigma = \sigma_m^2\tilde{\Sigma})$. They were assumed as a temporally independent with vector zero-mean and a covariance function matrix Σ specified by a Gaussian process. In this distribution, σ_m^2 denote the variance between sites for each pollutant p. The dense $S \times S$ correlation matrix ($\tilde{\Sigma}$) is given by the Matérn function, which depends on the Euclidean spatial distance. In this paper, we used a popular special case of Matérn family of covariance functions $C(d) = \exp(-\phi_p d)$, where d represents the distance between sites (km) and $\phi_p > 0$ describes the correlation strength (Cressie, 1993; Lindgren and Rue, 2015; Rao et al., 2012).

2.4. Methodology

To propose a compositional spatio-temporal air pollution model using a dynamic linear modelling framework, our approach encompasses the

following steps: (i) pre-processing air pollution data to express them as 6-part compositions, (ii) treating the missing data in the compositional data by imputation techniques, (iii) applying the DLM model to data, and (iv) evaluating the model in both fit and validation stages.

Both the proposed and conventional approaches were done with the software R studio (Figure 2). In R studio, the packages R2OpenBUGS (Sturtz et al., 2005), Compositions (Gerald van den Boogaart et al., 2018), and Openair (Carslaw and Ropkins, 2012) were used. The R script is described in Sánchez-Balseca and Pérez-Foguet (2020b).

2.4.1. Proposed approach

Step 1. Pre-processing of air pollution data. The daily pollutant volume concentrations were converted to concentration in weight (as part per billion, ppb), in order to use the compositional data approach. The daily air density at each monitoring station was used to convert the concentration units. The ideal gas law was used to determine air density. Eq. (4) shows air density (δ_{air}), which is calculated using temperature (T), pressure (P), and ideal gas constant for dry air ($R = 287.05\ \mathrm{Jkg^{-1}K^{-1}}$)

$$\delta_{air} = \frac{P}{R \cdot T}. \tag{8}$$

The closed composition can then be defined as $[SO_2, CO, O_3, NO_2, PM_{2.5}, Res]$, where *Res* is the residual or complementary part. We fixed $K = 1$ *billion* (as the concentration units used here were *ppb*). Considering the $sum(x) < K$ for all compositions x, we defined a complementary part as $Res = K - sum(x)$ for each day.

Step 2. Data processing. For missing daily data, we used the compositional robust imputation method: k-Nearest Neighbor Imputation (Martín-Fernández et al., 2003; Hron et al., 2008). This method requires at least one element in the composition. However, it is common to have days without information; for these timepoints, we chose the pollutant with less gaps to apply a simple imputation, with a season mean value, and then used the robust imputation method. To complement this stage, we presented a descriptive air pollution data analysis using the compositional biplot (Aitchison and Greenacre, 2002).

Step 3. Model application. To evaluate the compositional nature of data, two statistical approaches were used: (i) the conventional approach, where the 5 components (air pollutants) are modelled separately; and (ii) the compositional approach, in which the 6 components are first log-transformed into 5 coordinates (y_1^*, y_2^* , y_3^* , y_4^*, and y_5^*) using the method to build an optimal orthonormal basis (Gerald van den Boogaart et al., 2018). The log-ratio coordinates were then modelled separately, and finally regression results were back-transformed (see Section 2.3.1). However, to complement the proposed approach, it is necessary recover the original units for the estimates in compositional data analysis (Martín-Fernández et al., 2019). Once results are back-transformed in proportions (p^*_E; $sum(p^*_E) = 1$), they are multiplied by K to obtain the model results in original units.

Step 4. Model evaluation. The last step was the model evaluation, which comprised two stages: the first stage evaluated the fitted model, and the second one evaluated the model validation. For this, the Nash-Sutcliffe efficiency index (NSE), the relative squared error (RSE), and the Pearson correlation coefficient were used.

NSE (Eq. 5) is a widely used and potentially reliable statistic for assessing the goodness of fit of models. The NSE scale ranges from 0 to 1, whereby NSE = 1 means the model is perfect, NSE = 0 means that the model is equal to the average of the observed data, and negative values mean that the average is a better predictor (McCuen et al., 2006).

$$\mathrm{NSE} = 1 - \frac{\sum (\mathrm{Yobs_i} - \mathrm{Ysim_i})^2}{\sum (\mathrm{Yobs_i} - \overline{\mathrm{Yobs}})^2}. \tag{9}$$

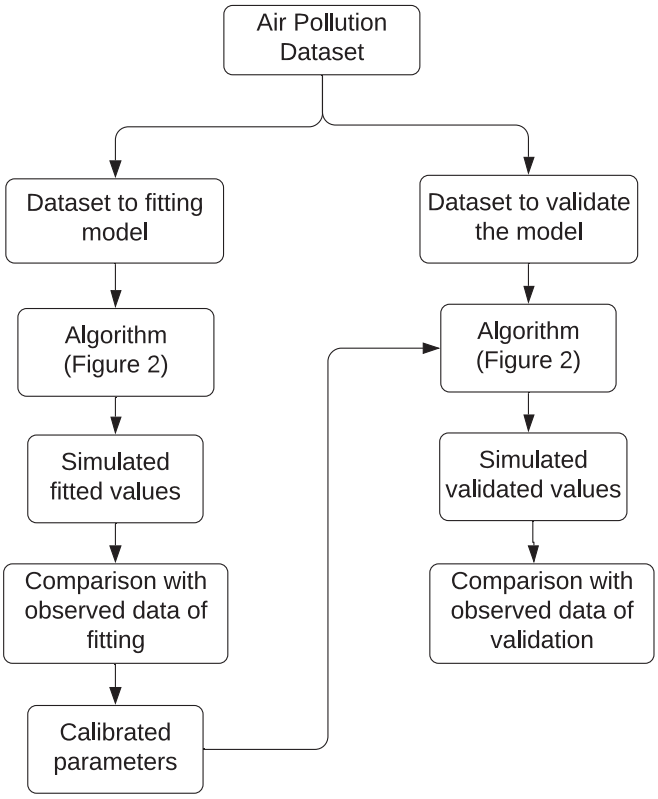

Figure 3. Scheme of fitting and validating modelling process.

RSE is a measure of quantitative performance commonly used to evaluate regression models (Eq. 6). A null RSE means that the model adjusts perfectly to comparison data. This criterion allowed compare between conventional and compositional approaches.

$$\mathrm{RSE} = \frac{\sum_{i=1}^{n} (\mathrm{Ysim_i} - \mathrm{Yobs_i})^2}{\sum_{i=1}^{n} (\overline{\mathrm{Yobs}} - \mathrm{Yobs_i})^2}. \tag{10}$$

The Pearson correlation coefficient is a linear measure between two quantitative random variables. Unlike RMSE, the NSE and Pearson correlation are independent of the scale of measurement of the variables. This coefficient can be calculated as shown in Eq. (7)

$$r_{(\mathrm{Yobs_i}, \mathrm{Ysim_i})} = \frac{n \sum \mathrm{Yobs_i} \cdot \mathrm{Ysim_i} - \sum \mathrm{Yobs_i} \sum \mathrm{Ysim_i}}{n \sum \mathrm{Yobs_i^2} - (\sum \mathrm{Yobs_i})^2 \sqrt{n \sum \mathrm{Ysim_i^2} - (\sum \mathrm{Ysim_i})^2}}. \tag{11}$$

2.4.2. Fitting and validating modelling

To fit the model, we used the first five years (2009–2013), and the last one (2014) for validation. Figure 3 summarizes the fitting and validation strategy. In the proposed model with the compositional approach, $\mathbf{y_{spt}}$ denote the Log-ratio coordinate at specific site ($s = 1, \ldots, 5$) on day t ($t = 1, \ldots, 1827$). The Bayesian approach through MCMC method was used to determine the model parameters ($\theta_{spt}, \sigma^2_{vsp}, \sigma^2_{wp}, \sigma^2_{mp}, m_{sp}, \phi_p$), and with them, we found the simulated log-ratio coordinates at the five monitoring stations for each day t ($t = 1, \ldots, 1827$). Finally, we evaluated the quality of the fitted model.

To validate the model, we used the model parameters obtained from Bayesian inference and the dataset of validation (Shaddick and Wakefield, 2002; Petris et al., 2007). Then, we applied the proposed algorithm (Figure 2) with the fixed model parameter from the fitting model process to the dataset of validation, and finally we evaluated the quality of evaluation stage.

$Yobs_i$ denote the observed five pollutant concentrations at five monitoring stations for both calibration and validation processes. $Ysim_i$

Table 3. Summary of daily data measured at five stations (2009–2013).

Variable	Unit.	Total	Missing	%	Mean	Min.	25%	Med.	75%	Max
ST_1										
CO	$mg{\cdot}m^{-3}$	1827	23	1.3	0.56	0.09	0.44	0.53	0.65	1.68
SO_2	$\mu g{\cdot}m^{-3}$	1827	14	0.8	3.24	0.12	1.89	2.87	4.19	12.9
O_3	$\mu g{\cdot}m^{-3}$	1827	13	0.7	26.9	1.52	21.9	25.8	30.5	68.3
NO_2	$\mu g{\cdot}m^{-3}$	1827	33	1.8	16.9	0.77	12.3	15.9	20.6	64.7
$PM_{2.5}$	$\mu g{\cdot}m^{-3}$	1827	48	2.6	19.6	5.67	13.9	18.1	23.1	99.5
ST_2										
CO	$mg{\cdot}m^{-3}$	1827	6	0.3	0.81	0.21	0.64	0.78	0.94	2.07
SO_2	$\mu g{\cdot}m^{-3}$	1827	6	0.3	5.06	0.23	3.25	4.80	6.39	20.3
O_3	$\mu g{\cdot}m^{-3}$	1827	5	0.3	20.9	2.47	14.4	19.06	24.5	78.2
NO_2	$\mu g{\cdot}m^{-3}$	1827	12	0.7	27.7	7.44	22.8	27.23	32.3	103.7
$PM_{2.5}$	$\mu g{\cdot}m^{-3}$	1827	22	1.2	17.3	3.86	13.5	17.12	20.9	57.3
ST_3										
CO	$mg{\cdot}m^{-3}$	1827	7	0.4	0.81	0.13	0.66	0.79	0.94	1.99
SO_2	$\mu g{\cdot}m^{-3}$	1827	12	0.7	8.11	0.58	4.14	6.09	9.76	56.6
O_3	$\mu g{\cdot}m^{-3}$	1827	6	0.3	23.1	6.22	17.0	21.4	26.5	84.4
NO_2	$\mu g{\cdot}m^{-3}$	1827	9	0.5	30.4	9.69	25.6	29.9	35.1	59.6
$PM_{2.5}$	$\mu g{\cdot}m^{-3}$	1827	17	0.9	21.9	7.74	16.9	21.0	26.4	124.3
ST_4										
CO	$mg{\cdot}m^{-3}$	1827	20	1.1	0.58	0.14	0.47	0.57	0.67	1.53
SO_2	$\mu g{\cdot}m^{-3}$	1827	25	1.4	3.48	0.24	2.35	3.23	4.35	18.8
O_3	$\mu g{\cdot}m^{-3}$	1827	7	0.4	22.9	1.35	16.5	21.3	27.5	78.8
NO_2	$\mu g{\cdot}m^{-3}$	1827	7	0.4	20.6	7.46	16.6	19.9	24.1	45.4
$PM_{2.5}$	$\mu g{\cdot}m^{-3}$	1827	22	1.2	16.2	1.24	12.4	15.7	19.2	89.8
ST_5										
CO	$mg{\cdot}m^{-3}$	1827	16	0.9	0.82	0.11	0.68	0.81	0.95	1.76
SO_2	$\mu g{\cdot}m^{-3}$	1827	21	1.1	4.61	0.13	3.03	4.19	5.77	18.6
O_3	$\mu g{\cdot}m^{-3}$	1827	9	0.5	23.6	0.71	17.0	21.5	27.8	76.5
NO_2	$\mu g{\cdot}m^{-3}$	1827	10	0.5	28.7	10.7	23.6	28.3	33.1	61.8
$PM_{2.5}$	$\mu g{\cdot}m^{-3}$	1827	2	0.1	18.6	0.87	14.3	18.0	22.5	86.0

denotes simulated five pollutant concentrations at five monitoring stations for both the calibration and validation processes. The quality metrics for the general model and for each monitoring station were obtained.

In this paper, covariate terms ($\boldsymbol{\beta}$, $\boldsymbol{X_{spt}}$) were not considered for clarity. A recent work proposes the use of covariates with an univariate model for particulate matter using satellite data in a wildfire event in Quito, showing the potential of the modelling approach (Sánchez-Balseca and Pérez-Foguet (2020a)).

3. Results

3.1. Pre-processing of air pollution data

Table 3 shows the summary of mean daily data calculated from the pre-processing data step. In general, the station ST_3 has higher air pollutants concentrations than the remain stations. The station ST_3 is located in a commercial zone with high vehicle mobility (Sánchez-Balseca, 2017; Secretaria de Ambiente DMQ, 2017). Other

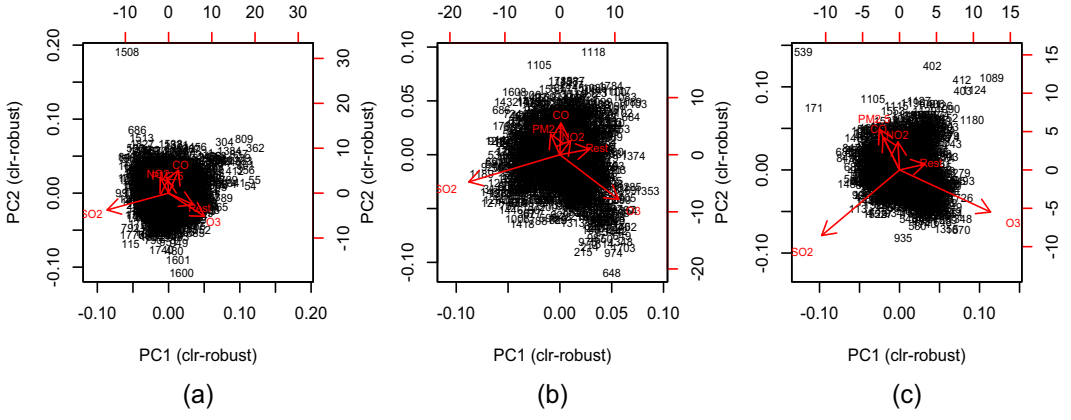

Figure 4. Biplots of *clr*-pollutants for (a) ST_1, (b) ST_3, and (c) ST_5.

Table 4. Posterior estimates (mean and first and last quantiles).

Parameter	Mean	2.5%	97.5%	Parameter	Mean	2.5%	97.5%
y_1^*							
σ_{v11}	0.1382	0.1255	0.1501	m_{11}	0.0450	–0.1067	0.1207
σ_{v21}	0.0591	0.0458	0.0712	m_{21}	0.1372	0.0809	0.2004
σ_{v31}	0.0941	0.0819	0.106	m_{31}	0.0957	–0.0073	0.2105
σ_{v41}	0.0725	0.0578	0.0869	m_{41}	0.0930	0.0281	0.1937
σ_{v51}	0.0577	0.0453	0.0695	m_{51}	0.1401	0.0747	0.2057
σ_{w1}	0.1645	0.1603	0.1687	ϕ_1	0.0524	0.0070	0.1107
σ_{m1}	0.1136	0.0585	0.24				
y_2^*							
σ_{v12}	0.0650	0.0474	0.0841	m_{12}	–0.05157	–0.1293	0.0304
σ_{v22}	0.1494	0.1336	0.1644	m_{22}	–0.0251	–0.1368	0.0855
σ_{v32}	0.1307	0.1144	0.1471	m_{32}	–0.06803	–0.1721	0.0904
σ_{v42}	0.1028	0.0872	0.1198	m_{42}	–0.03773	–0.1893	0.1313
σ_{v52}	0.1368	0.1216	0.1522	m_{52}	–0.07084	–0.1852	0.0656
σ_{w2}	0.2108	0.2054	0.2163	ϕ_2	0.06075	0.0079	0.1121
σ_{m2}	0.1094	0.0512	0.2298				
y_3^*							
σ_{v13}	0.1264	0.1162	0.1358	m_{13}	0.192	0.0885	0.292
σ_{v23}	0.0502	0.0392	0.0593	m_{23}	–0.0202	–0.0860	0.1109
σ_{v33}	0.0616	0.0519	0.0709	m_{33}	0.1087	0.0514	0.16
σ_{v43}	0.0644	0.0539	0.0746	m_{43}	0.1441	0.0246	0.2347
σ_{v53}	0.0519	0.0417	0.0622	m_{53}	0.1389	0.0815	0.1941
σ_{w3}	0.0939	0.0901	0.0976	ϕ_3	0.0698	0.0138	0.1131
σ_{m3}	0.1715	0.0788	0.3642				
y_4^*							
σ_{v14}	0.1218	0.109	0.1352	m_{14}	0.05996	–0.0419	0.166
σ_{v24}	0.0879	0.0741	0.1013	m_{24}	0.01439	–0.0669	0.0954
σ_{v34}	0.1028	0.0872	0.1198	m_{34}	–0.0495	–0.1522	0.0744
σ_{v44}	0.1292	0.1181	0.1411	m_{44}	0.03821	–0.0735	0.1182
σ_{v54}	0.1088	0.0979	0.12	m_{54}	–0.01232	–0.1096	0.0752
σ_{w4}	0.0986	0.0911	0.1046	ϕ_4	0.05479	0.0065	0.1118
σ_{m4}	0.0982	0.0484	0.2026				
y_5^*							
σ_{v15}	0.0698	0.0631	0.0764	m_{15}	0.04376	–0.0155	0.1164
σ_{v25}	0.0544	0.0479	0.0608	m_{25}	–0.05235	–0.1273	0.0019
σ_{v35}	0.0427	0.0363	0.0496	m_{35}	–0.0467	–0.0851	0.0114
σ_{v45}	0.0556	0.0489	0.0619	m_{45}	–0.06006	–0.136	0.0144
σ_{v55}	0.0563	0.05	0.0631	m_{55}	–0.2502	–0.0673	0.024
σ_{w5}	0.0977	0.0953	0.1	ϕ_5	0.06351	0.0098	0.1127
σ_{m5}	0.1015	0.0510	0.213				

areas with significant commercial activities are monitored by stations ST_2 and ST_5, which thus have higher SO_2 levels than the other zones (SO_2 is present in motor vehicle emissions, as the result of fuel combustion). Finally, ST_1 and ST_4 are located in residential zones. In generally, the station ST_1 has more missing values than the other stations. The missing data did not exceed about 2.6% in this work ($PM_{2.5}$).

3.2. Data processing

As ozone presented the lowest percentage of missing data (less than 0.7%) at all monitoring stations (see Table 3), it was initially used to apply the robust imputations descripted in the processing data step to all compositions.

The air quality data using the compositional approach followed some patterns in all monitoring stations. Figure 4 shows the biplot analysis for the odd-numbered monitoring stations (ST_1, ST_3, and ST_5); notably, the *clr*(SO_2), *clr*(O_3) variables had opposite directions than the others (in other words, an observation with high value of *clr*(SO_2) has a low value on the others coordinates, and vice-versa).

The variables *clr*(CO), *crl*(NO_2), and *clr*($PM_{2.5}$) had similar direction. This means that these variables are quite strong relations. The compositional biplot allowed us to identify the chemical behavior between them. For example, SO_2 also leads to photochemical O_3 production when CO, NO_2 and $PM_{2.5}$ are in low concentrations (Yoo et al., 2015).

3.3. Model fitting

Following with the model application step, the log-ratio coordinates were defined using the Table 2, where $x_1 = SO_2$, $x_2 = CO$, $x_3 = O_3$, $x_4 = NO_2$, $x_5 = PM_{2.5}$, and $x_6 = Res$.

Table 4 shows the principal posterior estimates (mean, first quantile, and fifth quantiles) of the model parameters σ_{vsp}, σ_{wp}, σ_{mp}, m_{sp}, and ϕ_p, where s and p denote the location and log-ratio coordinates, respectively. In agreement with the fact that the ST_1 station had more missing values than the remain stations (see Table 3), ST_1 had the most important measurement error variance (σ^2_{vsp}). In the first and second log-ratio coordinates, the temporal error variance (σ^2_{wp}) had more influence

Table 5. Fitted model evaluation criteria for both conventional and compositional approaches by air pollutants.

Metrics	CO		SO_2		O_3		NO_2		$PM_{2.5}$	
	Con.	Comp.	Con.	Comp.	Con.	Comp.	Con.	Comp.	Con.	Comp.
RSE	0.0758	0.0838	0.4362	0.0067	0.0191	0.0626	0.0603	0.0549	0.2369	0.0441
NSE	0.9242	0.9157	0.5637	0.9931	0.9809	0.9371	0.9397	0.9447	0.7628	0.9557
COR.	0.9630	0.9611	0.7557	0.9966	0.9918	0.9686	0.9702	0.9721	0.8768	0.9809

Comp., compositional approach; Con., conventional approach; Cor., correlation coefficient; NSE, Nash-Sutcliffe efficiency index; RSE, relative squared error.

Table 6. Fitted model evaluation criteria for both conventional and compositional approaches, by station and by pollutant.

ST	Metrics	CO		SO_2		O_3		NO_2		$PM_{2.5}$	
		Con.	Comp.	Con.	Comp.	Con.	Comp.	Con.	Comp.	Con.	Comp.
1	RSE	0.029	0.215	0.253	0.0448	0.023	0.1993	0.091	0.1277	0.256	0.0485
	NSE	0.971	0.812	0.747	0.9548	0.977	0.7941	0.909	0.8696	0.743	0.9509
	COR.	0.987	0.904	0.869	0.9943	0.990	0.9813	0.957	0.9480	0.869	0.9825
2	RSE	0.015	0.063	0.231	0.008	0.026	0.0741	0.124	0.1564	0.156	0.0282
	NSE	0.985	0.946	0.769	0.9915	0.974	0.9257	0.876	0.8429	0.844	0.9717
	COR.	0.993	0.974	0.881	0.9973	0.989	0.9756	0.939	0.9820	0.919	0.9890
3	RSE	0.100	0.072	0.669	0.0055	0.016	0.0331	0.076	0.0389	0.283	0.0434
	NSE	0.899	0.927	0.331	0.9944	0.984	0.9668	0.924	0.9609	0.717	0.9564
	COR.	0.979	0.970	0.615	0.9981	0.993	0.9842	0.963	0.9874	0.852	0.9894
4	RSE	0.421	0.052	0.255	0.0055	0.011	0.0243	0.078	0.0694	0.357	0.1293
	NSE	0.578	0.947	0.745	0.9944	0.988	0.9756	0.922	0.9299	0.643	0.8698
	COR.	0.998	0.976	0.867	0.9977	0.995	0.9892	0.962	0.9814	0.803	0.9434
5	RSE	0.077	0.051	3.669	0.0109	0.024	0.0496	0.093	0.0449	0.215	0.0515
	NSE	0.923	0.948	0.735	0.9889	0.976	0.9503	0.907	0.9548	0.785	0.9278
	COR.	0.992	0.979	0.859	0.9971	0.989	0.9801	0.954	0.9806	0.889	0.9735

Comp., compositional approach; Con., conventional approach; Cor., correlation coefficient; NSE, Nash-Sutcliffe efficiency index; RSE, relative squared error.

than spatial (σ^2_{mp}) and measurement error variance (σ^2_{vsp}). Thus, the influence of SO_2, CO, and O_3 over the composition had a temporal character that it was reflected through the temporal error variance defined by the system equation. The NO_2 influence over the composition (third log-ratio coordinate) is reflected by the high spatial error variance defined by Matérn function. In general, the influence of $PM_{2.5}$ (fourth log-ratio coordinate) over the composition is reflected by the high measurement error variance value in the stations ST_1, ST_3, ST_4, and ST_5. Note that the $PM_{2.5}$ dataset had more missing data than the other components (Table 3).

The ϕ_p mean model parameter allows the empirically derived correlation range (d_ϕ) to be calculated; that is, the distance at which the correlation is close to 0.05 (Sahu, 2012). This distance is a result of the Matérn function, which depends on the Euclidean spatial distance. The empirically derived correlation ranges for each log-ratio coordinate were 38.17, 32.92, 28.65, 36.50, and 31.49 km, respectively. The first and fourth log-ratio coordinates (the influence of SO_2, CO, and $PM_{2.5}$ over the composition) had correlations ranges that decreased slowly with distance. Taking into account the study case, all of correlations ranges were enough to cover a local territory where there are limited monitoring stations. The third log-ratio coordinate (the NO_2 influence over the composition) had higher values of site effects (m_{sp}) than the others coordinates; thus, this correlation range value is relatively low (28.65 km). The posterior of ϕ_p values were very diffuse due to the small number of monitoring stations at urban area (Shaddick and Wakefield, 2002).

The correlation range for the fourth log-ratio coordinate that represents the influence of $PM_{2.5}$ over the composition ($d_\phi = 36.68$) is slightly greater than the correlation range of $PM_{2.5}$ in the conventional model for the univariate approach ($d_\phi = 27.19$) (Sánchez-Balseca and Pérez-Foguet (2020a)).

Table 5 shows the evaluation criteria values of the fitted model for both the conventional and the compositional approach. In general, both approaches presented good quality modelling (NSE > 0.5), and adequate fitted data for all pollutants (RSE close to zero) (Ritter and Muñoz-Carpena, 2013). For CO and NO_2, the behavior was similar between both model approaches. For SO_2 and $PM_{2.5}$, the compositional model approach presented a better model quality and fitted values than the conventional approach. For O_3, the conventional approach gave values slightly better than the compositional approach.

Table 6 presents the results by monitoring stations. Main result is that the conventional model at monitoring station ST_3 had values of NSE <0.5 for SO_2. The SO_2 values at all stations (but mainly at ST_3) have more variability than the values of the others pollutants (Figure 4). The high variability of SO_2 is due to the location of monitoring stations near to the paths in the urban area, and they can recorder levels of anthropogenic SO_2, such as vehicle emissions. Figure 5 shows the observed data and fitted data of the five pollutants at monitoring stations ST_1, ST_3, and ST_5 using both conventional and compositional approaches.

3.4. Model validation

The compositional approach presented lower RSE values than the conventional approach and gave positive NSE values for all pollutants. The correlation coefficient for CO, O_3, and NO_2 presented an adequate correlation with the observed time series. The quality metrics for the validation process of the conventional approach are shown in Table 7.

Table 8 shows the quality metrics by station and pollutant. The negative NSE values in both models by pollutant and station indicate that the mean value of the observed data was better for validating the model. For example, the simulated values of SO_2 at ST_1, ST_3, and ST_5 in Figure 6 follow a horizontal trend (constant mean value) with little

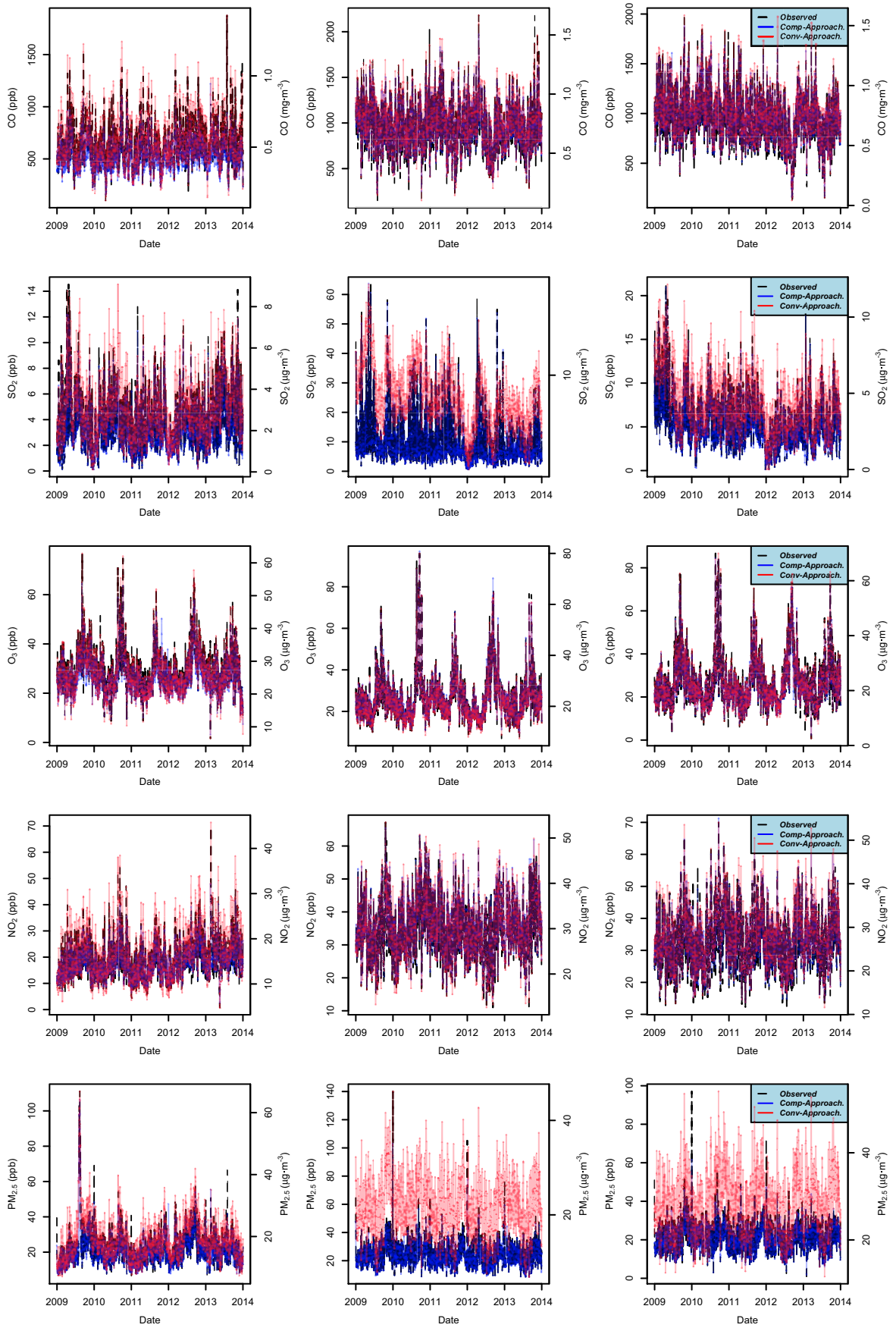

Figure 5. Fitted model for each pollutant using both conventional (red curve) and compositional (blue curve) approaches at monitoring stations: ST_1(left), ST_3 (center), and ST_5 (right). The observed values are represented by the black curve.

Table 7. Validated model evaluation criteria for both conventional and compositional approaches by air pollutants.

Metrics	CO		SO_2		O_3		NO_2		$PM_{2.5}$	
	Con.	Comp.	Con.	Comp.	Con.	Comp.	Con.	Comp.	Con.	Comp.
RSE	0.9699	0.8761	2.5184	0.9246	0.6552	0.6930	0.6414	0.6806	1.0076	0.8919
NSE	-0.003	0.0944	-1.518	0.0675	0.3318	0.2932	0.3339	0.2940	-0.037	0.0814
COR.	0.3540	0.5293	-0.023	0.3773	0.5979	0.5593	0.5997	0.5952	0.2303	0.3589

Comp., compositional approach; Con., conventional approach; Cor., correlation coefficient; NSE, Nash-Sutcliffe efficiency index; RSE, relative squared error.

Table 8. Validated model evaluation criteria for both conventional and compositional approaches by station and pollutant.

ST	Metrics	CO		SO_2		O_3		NO_2		$PM_{2.5}$	
		Con.	Comp.	Con.	Comp.	Con.	Comp.	Con.	Comp.	Con.	Comp.
1	RSE	1.2065	1.0633	1.1980	1.4369	0.6721	0.8076	1.1737	1.4335	1.7266	1.1086
	NSE	-0.206	-0.063	-0.198	-0.437	0.3276	0.1921	-0.194	-0.459	-0.730	-0.111
	COR.	0.0034	0.3073	-0.254	0.0581	0.5136	0.5672	0.0453	0.3592	-0.257	0.2446
2	RSE	1.5381	2.0561	1.1539	1.2210	0.6906	0.7370	0.8393	0.9993	0.9424	0.9409
	NSE	-0.538	-1.056	-0.153	-0.221	0.3094	0.2630	0.1607	0.0007	0.0575	0.0591
	COR.	0.1711	0.1601	0.1205	0.1014	0.5719	0.5648	0.4080	0.3775	0.3262	0.3336
3	RSE	0.5698	0.5344	0.8165	0.9634	0.5363	0.5118	0.4714	0.4688	0.5965	0.5321
	NSE	-0.110	-0.046	-0.020	-0.205	0.2161	0.2503	0.0648	0.0670	0.0528	0.1524
	COR.	0.1920	0.2115	0.1921	0.0499	0.5582	0.5583	0.2747	0.2897	0.2463	0.3980
4	RSE	2.4135	1.3198	1.1278	0.2500	0.5844	0.6069	0.7717	0.7718	1.1453	1.2153
	NSE	-1.413	-0.319	-0.127	-0.132	0.4155	0.3930	0.2282	0.2281	-0.145	-0.215
	COR.	-0.154	-0.151	0.1193	0.0806	0.6551	0.6595	0.4834	0.4817	0.1082	0.1593
5	RSE	1.0118	1.2181	1.2441	1.2765	0.7447	0.7175	0.9253	0.9615	1.0300	0.9920
	NSE	-0.011	-0.128	-0.244	-0.276	0.2553	0.2824	0.0747	0.0385	-0.031	0.0067
	COR.	0.2798	0.2977	-0.074	-0.099	0.5631	0.5631	0.3352	0.3229	0.2071	0.2617

Comp., compositional approach; Con., conventional approach; Cor., correlation coefficient; NSE, Nash-Sutcliffe efficiency index; RSE, relative squared error.

variation; in this case, it is better to use the mean value to validate the model.

4. Discussion

The present article evaluated a spatio-temporal air pollution model with dynamic linear framework using five pollutants concentrations at five monitoring stations to determine the effectiveness of compositional approach. It is important to note that our methodology complements the conventional air pollution modelling in order to improve it, but that it was necessary to make the comparison between both approaches. Although our proposed model with Bayesian inference allows us to calculate missing observations, the missing data percentage affected more the multivariate analysis in the conventional modelling. This behavior was explained by Martín-Fernández and Thió-Henestrosa (2006) through the distance used in the imputation process. The classical statistical treatment of missing multivariate data uses the Euclidean distance measure, and the compositional approach uses the Aitchison distance. The Aitchison distance highlights the property that indicates that composition data have information only in the relationships between the parts (Hron et al., 2008). The information in the coordinates allows that the compositional approach had better quality metrics than the conventional approach. However, the main disadvantage of our proposed methodology was the difficulty to interpret the log-ratio coordinates and their related parameters. It can helpful to understand the log-ratio coordinates as the influence of one pollutant over the composition.

Several authors have previously presented works to determine the relationship between air pollutants through different univariate models (one to one) (Gimeno et al., 1997; Yoo et al., 2015; Xiao et al., 2018). In our approach, however, the biplot analysis allows a fast and adequate multivariate analysis between all available pollutants. This analysis considers the compositional character in the air pollution dataset. Recent contributions allow covariates and their relationship with the environmental data to be added (Daunis-i-Estadella et al., 2011). The compositional modelling presented better model evaluation criteria than the conventional approach for pollutants with high concentration and variability (e.g., SO_2). We will explain this result with a simple example. Consider under a conventional approach a composition of two air pollutants $[p_1, p_2]$ with a sum of one billion parts, if the pollutant p_1 increases its concentration, the pollutant concentration of p_2 must be reduced, and vice versa. The two pollutants have an inverse correlation imposed upon them, even if these two pollutants have no relationship. This imposed inverse correlation is called "spurious correlation" (Filzmoser et al., 2010; Egozcue et al., 2012), and log-ratio coordinates have been proposed to solve it (Aitchison, 1982; Egozcue et al., 2003). The simple example described could be used for one air pollutant using our method, where p_2 could be the residual or complementary part (Filzmoser et al., 2010).

The main limitation of our proposed model is that it used only one temporal covariate (first-order autoregressive variable). We used only one temporal covariate with the purpose of clarifying the application of compositional data in space–time models in air quality. As shown in the previous section, this consideration does not allow adequate quality metrics to be obtained in the validation process for either the conventional or the compositional approach. Further work with the proposed approach could add meteorological covariates. For instance, an univariate dynamic linear model was used for particulate matter in wildfire scenarios in Quito (Sánchez-Balseca and Pérez-Foguet (2020a)). Additional investigations could evaluate other spatial approaches, such as the Gaussian Markov random field through the stochastic partial differential equations.

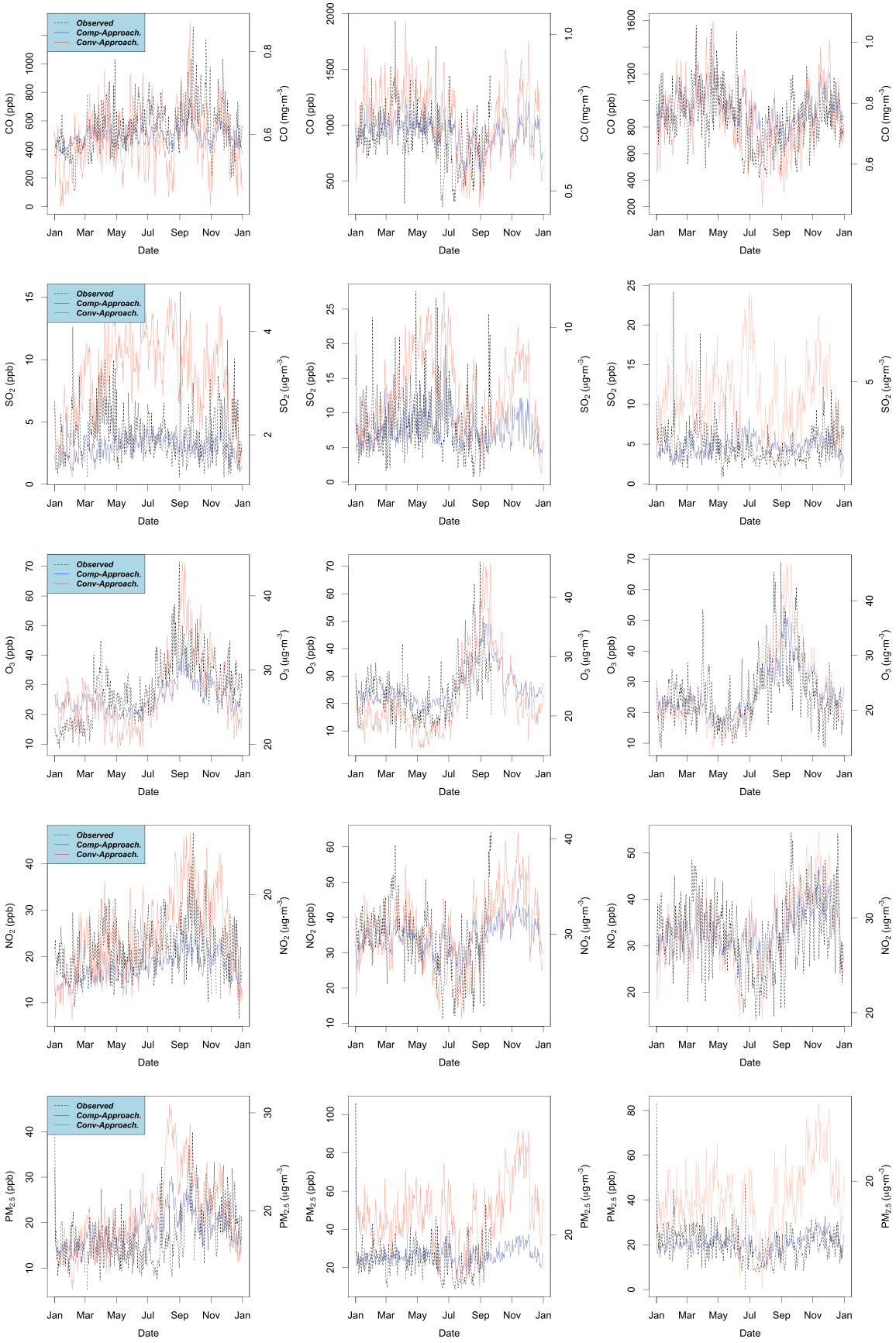

Figure 6. Model validation for each pollutant using both a conventional approach in 2014 (red curve) and a compositional (blue curve) approach at the monitoring stations ST_1(left), ST_3 (center), and ST_5 (right). The observed values are represented by the black curve.

5. Conclusion

The compositional approach used here improved the classical spatio-temporal air pollution modelling based on dynamic linear models. The robust imputation method using the compositional approach allowed the proportion between air pollution data to be maintained, thereby avoiding spurious correlations between them. The compositional biplot analysis allows a fast and adequate overview to the chemical relationships between atmospheric pollutants. This proposed model enhances dynamic linear modelling of pollutants with high variability (e.g., SO_2). Further, compositional modelling improves spatial-correlation descriptions of standard non-compositional approaches. Finally, this novel approach allows better predictions to be made about the levels of pollutants over an urban territory at time *t*.

Declarations

Author contribution statement

Joseph Sánchez-Balseca: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Agustí Pérez-Foguet: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

Data related to this article can be found at http://190.11.24.212/reportes/ReporteHorariosData.aspx, an open-source online data repository hosted at Secretaria de Ambiente del Distrito Metropolitano de Quito (Secretaria de Ambiente DMQ, 2015).

Acknowledgements

Joseph Sánchez Balseca is the recipient of a full scholarship from the Secretaria de Educación Superior, Ciencia, Técnología e Innovación (SENESCYT), Ecuador. The authors want to thank the CoDa knowledge management to the Ministry of Science, Innovation and Universities of Spain (Ref: RTI2018-095518-B-C22) and the Agència de Gestió d'Ajuts Universitaris i de Recerca de la Generalitat de Catalunya (Ref. 2017 SGR 1496).

References

Aitchison, J., 1982. The statistical analysis of compositional data (with discussion). J. Roy. Stat. Soc. 44 (2), 139–177.
Aitchison, J., Greenacre, M., 2002. Biplots for compositional data. J. Roy. Stat. Soc. 51 (4), 375–392.
AL-Dhurafi, N.A., Masseran, N., Zamzuri, Z., 2018. Compositional time series analysis for Air Pollution Index data. Stoch. Environ. Res. Risk Assess. 32 (10), 2903–2911.
Arakia, S., Shima, M., Yamamoto, K., 2018. Spatiotemporal land use random forest model for estimating metropolitan NO2 exposure in Japan. Sci. Total Environ. 634, 1269–1277.
Arroyo, V., Linares, C., Díaz, J., 2019. Premature births in Spain: measuring the impact of air pollution using time series analyses. Sci. Total Environ. 660, 105–114.
Blake, S., Henry, T., Murray, J., Flood, R., Muller, M.R., Jones, A.G., Rath, V., 2016. Compositional multivariate statistical analysis of thermal groundwater provenance: a hydrogeochemical case study from Ireland. Appl. Geochem. 75, 171–188.
Ban, Y., Zhang, P., Nascetti, A., Bevington, A., Wulder, M., 2020. Near real-time wildfire progression monitoring with Sentinel-1 SAR time series and deep learning. Sci. Rep. 10 (1322), 1–15.
Blangiardo, M., Pirani, M., Kanapka, L., Hansell, A., Fuller, G., 2019. A hierarchical modelling approach to assess multi pollutant effects in time-series studies. PloS One 14 (3), e0212565.
Buccianti, A., Pawlowsky-Glahn, V., 2005. New perspectives on water chemistry and compositional data analysis. Math. Geol. 37 (7), 703–727.
Buccianti, A., Mateu, G., Pawlowsky, V., 2006. Compositional Data Analysis in the Geosciences. Geological Society, London.
Cameletti, M., Ignaccolo, R., Bande, S., 2011. Comparing spatio-temporal models for particulate matter in Piemonte. Environmetrics 22 (8), 985–996.
Carslaw, D., Ropkins, K., 2012. Openair – an R package for air quality data analysis. Environ. Model. Software 27–28, 52–61.
Cetin, M., Sevik, H., Isinkaralar, K., 2017. Changes in the particulate matter and CO2 concentrations based on the time and weather conditions: the case of Kastamonu. Oxid. Commun. 40 (1), 477–485.
Cocchi, D., Greco, F., Trivisano, C., 2007. Hierarchical space-time modelling of PM10 pollution. Atmos. Environ. 41 (3), 532–542.
Cressie, N., 1993. Statistics for spatial data, Revised Ed. John Wiley & Sons, Michigan.
Daunis-i-Estadella, J., Thió-Henestrosa, S., Mateu-Figueras, G., 2011. Including supplementary elements in a compositional biplot. Comput. Geosci. 37 (5), 696–701.
Dominici, F., McDermott, A., Zeger, S., Samet, J., 2002. On the use of generalized additive models in time-series studies of air pollution and health. Am. J. Epidemiol. 156 (3), 193–203.
Egozcue, J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C., 2003. Isometric logratio transformations for compositional data analysis. Math. Geol. 35, 279–300.
Egozcue, J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Math. Geol. 37, 795–828.
Egozcue, J.J., Daunis-i-Estadella, J., Pawlowsky-Glahn, V., Hron, K., Filzmoser, P., 2012. Simplicial regression. The normal model. J. Appl. Probab. Stat. 6 (1), 87–108.
EMASEO, 2011. Plan de Desarrollo 2012-2022. Municipio del Distrito Metropolitano de Quito. Municipio del Distrito Metropolitano de Quito, Quito.
Environmental Protection Agency, 2001. AQS Report. Environmental Protection Agency. EPA, Washinton.
Ezbakhe, F., Pérez-Foguet, A., 2019. Estimating Access to drinking water and sanitation: the need to account for uncertainty in trend analysis. Sci. Total Environ. 696, 133830.
Fassò, A., Finazzi, F., 2011. Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. Environmentrics 22 (6), 735–748.
Filzmoser, P., Hron, K., Reimann, C., 2010. The bivariate statistical analysis of environmental (compositional) data. Sci. Total Environ. 408 (19), 4230–4238.
Gerald van den Boogaart, K., Tolosana-Delgado, R., Bren, M., 2018. Compositions: compositional data analysis. R package versión 1. 40–42. Available at. https://cran.r-project.org/web/packages/compositions.
Gimeno, L., Rua, A., Hernández, E., 1997. Relationship between air pollutants emission patterns and concentrations. Toxicol. Environ. Chem. 57 (1-4), 189–197.
Gutiérres, L., Mena, R., Ruggiero, M., 2016. A time dependent Bayesian nonparametric model for air quality analysis. Comput. Stat. Data Anal. 95, 161–175.
He, J., Ding, S., Liu, D., 2019. Exploring the spatiotemporal pattern of $PM_{2.5}$ distribution and its determinants in Chinese cities based on a multilevel analysis approach. Sci. Total Environ. 659, 1513–1525.
Hron, K., Templ, M., Filzmoser, P., 2008. Imputation of Compositional Data Using Robust Methods. Vienna University of Technology, Department of Statistics and Probability Theory. Vienna University of Technology, Vienna.
Hu, H., Hu, Z., Zhong, K., Xu, J., Zhang, F., Zhao, Y., Wu, P., 2019. Satellite-based high-resolution mapping of ground-level $PM_{2.5}$ concentrations over East China using a spatiotemporal regression kriging model. Sci. Total Environ. 672, 479–490.
Huang, Y., Yao, T., Fung, J.C., Lu, X., Lau, A.K., 2019. Application of air parcel residence time analysis for air pollution prevention and control policy in the Pearl River Delta region. Sci. Total Environ. 658, 744–752.
Jarauta-Bragulat, E., Hervada-Sala, C., Egozcue, J.J., 2016. Air quality index revisited from a compositional point of view. Math. Geosci. 48 (5), 581–593.
Jato-Espino, D., Castillo-Lopez, E., Rodriguez-Hernandez, J., Ballester Munoz, F., 2018. Air quality modelling in Catalonia from a combination of solar radiation, surface reflectance and elevation. Sci. Total Environ. 624, 189–200.
Jurado, J., Southgate, D., 1999. Dealing with air pollution in Latin America: thecase of Quito, Ecuador. Environ. Dev. Econ. 4 (3), 375–388.
Lindgren, F., Rue, H., 2015. Bayesian spatial modelling with R-INLA. J. Stat. Software 63 (19), 1–25.
López-Abente, G., Locutura-Rupérez, J., Fernández-Navarro, P., Martín-Méndez, I., Bel-Lan, A., Núñez, O., 2018. Compositional analysis of topsoil metals and its associations with cancer mortality using spatial misaligned data. Environ. Geochem. Health 40 (1), 283–294.
Marinov, M., Tapalov, I., Gieva, E., Nicolov, G., 2016. Air quality monitoring in urban environments. In: 39th International Spring Seminar on Electronics Technology (ISSE). IEEE, Pilsen, pp. 443–448.
Martín-Fernandez, J.A., Raju, N.J., Egozcue, J.J., Pawlowsky-Glahn, V., Olea, R.A., 2019. How to recover the original units for the estimates in compositional data analysis?. In: The 8th International Workshop on Compositional Data Analysis. 17. Terrassa: CODAWORK.
Martín-Fernández, J., Thió-Henestrosa, S., 2006. Rounded zeros: some practical aspects for compositional data. In: Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (Eds.), Compositional Data Analysis in the Geosciences (191-201). The Geological Society, London.
Martín-Fernández, J., Barceló-Vidal, C., Pawloswsky-Glahn, V., 2003. Dealing with zeros and missing values in compositional data sets using nonpara- metric imputation. Math. Geol. 35 (3), 253–278.
Mayer, H., 1999. Air pollution in cities. Atmos. Environ. 33 (24-25), 4029–4037.

McCuen, R.H., Knight, Z., Cutter, A.G., 2006. Evaluation of the nash–Sutcliffe efficiency index. J. Hydrol. Eng. 11 (6), 597–602.
Meagher, J.F., Lee, N.T., Parkhurst, W.J., 1967. Rural ozone in the southeastern United States. Atmos. Environ. 21 (3), 605–615.
Mölter, A., Lindley, S., de Vocht, F., Simpson, A., Agius, R., 2010. Modelling air pollution for epidemiologic research – Part II: predicting temporal variation through land use regression. Sci. Total Environ. 409 (1), 211–217.
Owen, D.D., Pawlowsky-Glahn, V., Buccianti, A., Bradd, J.M., 2016. Compositional data analysis as a robust tool to delineate hydrochemical facies within and between gas-bearing aquifers. Water Resour. Res. 52 (8), 5771–5793.
Paci, L., 2013. Bayesian Space-Time Data Fusion for Real-Time Forecasting and Map Uncertainty. Universit`a di Bologna, Bologna.
Pak, U., Ma, J., Ryu, U., Ryom, K., Juhyoke, U., Pak, K., Pak, C., 2020. Deep learning-based PM2.5 prediction considering the spatiotemporal correlations: a case study of Beijing, China. Sci. Total Environ. 699 (133561).
Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015. Modeling and Analysis of Compositional Data. John Wiley & Sons, London.
Pérez-Foguet, A., Giné-Garriga, R., Ortego, M., 2017. Compositional data for global monitoring: the case of drinking water and sanitation. Sci. Total Environ. 590–591, 554–565.
Petrika, A., Thiombanea, M., Lima, A., Albanese, S., Buscher, J.T., De Vivo, B., 2018. Soil contamination compositional index: a new approach to quantify contamination demonstrated by assessing compositional source patterns of potentially toxic elements in the Campania Region (Italy). Appl. Geochem. 96, 264–276.
Petris, G., Petrone, S., Campagnoli, P., 2007. Dynamic Linear Models with R. Springer, Berlin.
Quispe-Coica, A., Pérez-Foguet, A., 2020. Preprocessing alternatives for compositional data related to water, sanitation and higiene. Sci. Total Environ. in press.
Rao, T.S., Rao, S.S., Rao, C., 2012. Handbook of Statistics: Time Series Analysis, Vol. 30. Elsevier, Amsterdam.
Reimann, C., Filzmoser, P., Fabian, K., Hron, K., Birke, M., Demetriades, A., Ladenberger, A., 2012. The concept of compositional data analysis in practice — Total major element concentrations in agricultural and grazing land soils of Europe. Sci. Total Environ. 426, 196–210.
Reimann, C., Filzmoser, P., Hron, K., Kynclova, P., Garrett, R., 2017. A new method for correlation analysis of compositional (environmental) data – a worked example. Sci. Total Environ. 697–608, 965–971.
Ritter, A., Munoz Carpena, R., 2013. Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. J. Hydrol. 480, 33–45.
Sahu, S., 2012. Handbook of Statistics - Hierarchical Bayesian Models for Space–Time Air Pollution Data. Elsevier, Southampton.
Sánchez-Balseca, J., 2017. Indicadores socio ambientales para fortalecer la sostenibilidad de la política de movilidad en el GAD del Distrito Metropolitano de Quito. Caso: sector El Camal. Pontificia Universidad Católica del Ecuador, Quito.
Sánchez-Balseca, J., Pérez-Foguet, A., 2019. Assesing CoDa regression for modelling daily multivariate air pollutants evolution. In: 8th International Workshop on Compositional Data Analysis (CoDaWork2019). Universitat Politècnica de Catalunya-BarcelonaTECH, Terrassa, pp. 143–150.
Sánchez-Balseca, J., Pérez-Foguet, A., 2020a. Modelling hourly spatio-temporal PM2.5 concentration in wildfire scenarios using dynamic linear models. Atmos. Res. 242.
Sánchez-Balseca, J., Pérez-Foguet, A., 2020b. Spatio-temporal air pollution modelling using a compositional approach (dataset and R code). Zenodo.
Secretatia de Ambiente del DMQ, 2015. Reporte de calidad del aire, especies medidas. http://190.11.24.212/reportes/ReporteHorariosData.aspx.
Secretaria de Ambiente DMQ, 2017. Informe de la Calidad de Aire-2016 Distrito Metropolitano de Quito. Distrito Metropolitano de Quito, Quito.
Shaddick, G., Wakefield, J., 2002. Modelling daily multivariate pollutant data at multiple sites. J. Roy. Stat. Soc. 51 (3), 351–372. www.jstor.org/stable/3592657.
Shaddick, G., Thomas, M.L., Jobling, A., Brauer, M., Donkelaar, A.V., Burnett, R., Prüss-Ustün, A., 2018. Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. Roy. Stat. Soc. 231–253.
Shi, X., Zhao, C., Jiang, J.H., Wang, C., Yang, X., Yung, Y.L., 2018. Spatial representativeness of $PM_{2.5}$ concentrations obtained using observations from network stations. J. Geophys. Res. Atmos. 123 (6), 3145–3158.
Shi-wen, Z., Chong-yang, S., Xiao-yang, C., Hui-chun, Y., Yuan-fang, H., Shuang, L., 2013. Spatial interpolation of soil texture using compositional kriging and regression kriging with consideration of the characteristics of compositional data and environment variables. J. Integr. Agric. 12 (9), 1673–1683.
Sturtz, S., Ligges, U., Gelman, A., 2005. R2WinBUGS: a package for running WinBUGS from R. J. Stat. Software 12 (3), 1–16.
Vlachokostas, C., Achillas, C., Moussiopoulos, N., Hourdakis, E., Tsilingiridis, G., Ntziachristos, L., Sidiropoulos, C., 2009. Decision support system for the evaluation of urban air pollution control options: application for particulate pollution in Thessaloniki, Greece. Sci. Total Environ. 407 (23), 5937–5948.
Xiao, K., Wang, Y., Wu, G., Fu, B., Zhu, Y., 2018. Spatiotemporal characteristics of air pollutants (PM10, PM2.5, SO2, NO2, O3, and CO) in the inland basin city of chengdu, Southwest China. Atmosphere 9 (2), 74.
Yang, Y., Zhao, C., Sun, L., Wei, J., 2019. Improved aerosol retrievals over complex regions using NPP Visible Infrared Imaging Radiometer Suite observations. Earth Space . 6 (4), 629–645.
Yao, J., Brauer, M., Raffuse, S., Henderson, S., 2018. A machine learning approach to estimate hourly exposure to fine particulate matter for urban, rural, and remote populations during wildfire seasons. Environ. Sci. Technol. 52 (22), 13239–13249.
Yoo, J.-M., Jeong, M.-J., Kim, D., KStockwell, W.R., Yang, J.-H., Shin, H.-W., Lee, S.-D., 2015. Spatiotemporal variations of air pollutants O3, NO2, SO2, CO, PM10, and VOCs with land-use types. Atmos. Chem. Phys. 15 (18), 10857–10885.
Zannetti, P., 1990. Air Pollution Modeling. Springer Science, Monrovia, California.
Zhang, Y., 2019. Dynamic effect analysis of meteorological conditions on air pollution: a case study from Beijing. Sci. Total Environ. 684, 178–185.
Zhao, C., Wang, Y., Shi, X., Zhang, D., Wang, C., Jiang, J.H., Zhang, Q., Fan, H., 2019. Estimating the contribution of local primary emissions to particulate pollution using high-density station observations. J. Geophys. Res.: Atmospheres 124 (3), 1–14.
Zheng, C., Zhao, C., Zhu, Y., Wang, Y., Shi, X., Wu, X., Chen, T., Wu, F., Qiu, Y., 2017. Analysis of influential factors for the relationship between $PM_{2.5}$ and AOD in Beijing. Atmos. Chem. Phys. 17 (21), 13473–13489.
Zheng, C., Zhao, C., Li, Y., Wu, X., Zhang, K., Gao, J., Qiao, Q., Ren, Y., Zhang, X., Chai, F., 2018. Spatial and temporal distribution of NO_2 and SO_2 in Inner Mongolia urban agglomeration obtained from satellite remote sensing and ground observations. Atmos. Environ. 188, 50–59.