Environmental Research 204 (2022) 112388

Contents lists available at ScienceDirect

Environmental Research

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

Compositional and Bayesian inference analysis of the concentrations of air pollutants in Catalonia, Spain

Anna Mota-Bertran [a,b], Marc Saez [a,b,*], Germà Coenders [a,b]

[a] Research Group on Statistics, Econometrics and Health (GRECS), University of Girona, Girona, Spain
[b] CIBER of Epidemiology and Public Health (CIBERESP), Madrid, Spain

ARTICLE INFO

Keywords:
Compositional data
Integrated nested laplace approximation
Air pollution
Air pollutant
T-spaces

ABSTRACT

While most countries have networks of stations for monitoring pollutant concentrations, they do not cover the whole territory continuously. Therefore, to be able to carry out a spatial and temporal study, the predictions for air pollution from unmeasured sites and time periods need to be used.

The objective of this study is to predict the air pollutant concentrations of PM_{10}, O_3, NO_2, SO_2 and CO in sites throughout Catalonia (Spain) and time periods without a monitoring station. Compositional data (CoDa) studies the relative importance of pollutants. A novel feature in this article is combining CoDa with an indicator of total pollution. Predictions are then made using a combination of spatio-temporal models and the Bayesian Laplace Integrated Approach (INLA) inference method.

The most relevant results obtained indicate that the log-ratio between NO_2 and O_3 has the highest variance and the best predictive accuracy in time and space. Total pollution levels are second in variance but have low spatial predictive accuracy. Third in variance is the low temporal accuracy found in the log-ratio between SO_2 and the remaining pollutants. Globally, the combination of CoDa and the INLA method is suitable for making effective spatio-temporal predictions of air pollutants.

1. Introduction

Outdoor pollution caused around 249,000 premature deaths in 2016, and 83,000 of those deaths occurred as a result of the air pollution produced from the use of solid fuels in the home. What should also be borne in mind is that a close relationship exists between inequalities in development, non/compliance with environmental laws, regulations and policies and the exposure to pollution of different population groups (PAHO, 2017).

Air pollution has gained recognition and prominence on global agendas. In September 2015, the United Nations General Assembly adopted the 2030 Agenda for Sustainable Development. The central references to air pollution in the 2030 Agenda are to be found in Target 3.9 (Substantially reduce the number of deaths and illnesses caused by hazardous chemicals and air, water and soil pollution), Target 7.1 (Guarantee universal access to affordable, reliable and modern energy services) and Target 11.6 (Reduce the negative per capita environmental impact of cities, including paying special attention to air quality and municipal and other waste management) (WHO, 2018).

According to the European Environment Information and Observation Network, the main polluting gases and particles that most affect human health and the environment are coarse particles, (i.e., PM_{10} with a diameter of 10 μm (μm) or less), nitrogen dioxide (NO_2), ozone (O_3), carbon monoxide (CO) and sulphur dioxide (SO_2) (EIONET, 2020; Sicard et al., 2021).

Most countries have networks of stations for monitoring pollutant concentrations, but the main problem with these networks is that they do not cover the whole territory in a homogeneous way. Furthermore, the existing stations are not always in continuous operation which means they cannot provide data throughout the study period in question. Therefore, to be able to carry out a spatial and temporal study, the predictions for air pollution in sites and time periods without a monitoring station need to be used.

The scientific literature has provided a number of alternative methods and models to make this type of prediction. For instance, the hierarchical spatio-temporal models applied by Cameletti et al. focused on particulate matter (PM_{10}) in the Piemonte Region (Cameletti et al., 2011, 2013); the two-stage Bayesian model used by Blangiardo et al. to

* Corresponding author. Research Group on Statistics, Econometrics and Health (GRECS) and CIBER of Epidemiology and Public Health (CIBERESP), University of Girona, Carrer de la Universitat de Girona 10, Campus de Montilivi, 17003, Girona, Spain.
E-mail address: marc.saez@udg.edu (M. Saez).

https://doi.org/10.1016/j.envres.2021.112388
Received 26 July 2021; Received in revised form 12 November 2021; Accepted 12 November 2021
Available online 19 November 2021
0013-9351/© 2021 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

estimate the monthly concentration of NO_2 (Blangiardo et al., 2016); the Bayesian spatial-temporal analysis employed by Liu et al. (2020) to analyse the association between PM_{10}, SO_2 and NO_2 in the Chinese province of Hubei (Liu et al., 2020), the several machine learning methods compared by Liang et al. (2020) to predict the Taiwanese air quality index, and the super-Gaussian geometry methods used by Jia and Kikumoto (2021), to cite just a few of the many studies on the subject.

This article adds to the literature on prediction methods for air pollution in sites and time periods without a monitoring station by combining two novel approaches. On the one hand, it uses Bayesian inference with the Integrated Nested Laplace Approximation (INLA) (Rue et al., 2009, 2017; Lindgren and Rue, 2015) to take into account time, space, and covariates. On the other hand, it completes the classic approach of analysing relative concentrations of chemicals as Compositional Data-CoDa (Pawlowsky-Glahn et al., 2015a; Sánchez-Balseca and Pérez-Foguet, 2020), by adding a total pollution index by means of T-spaces (Pawlowsky-Glahn et al., 2015b). Compared to standard air pollution studies, CoDa makes it possible to study how the concentrations of the different pollutants can increase or decrease with respect to one another. Furthermore, the novel introduction of T-spaces makes the computation of a global air quality index compatible with the compositional perspective.

The structure of the article is as follows. First, we explain the compositional data approach to air pollutant concentrations. Second, we present a Bayesian approach; specifically, the INLA method to be able to make effective spatio-temporal predictions. Then, we make a predictive accuracy assessment and finally we discuss the results.

2. Methods

2.1. Design

The research was conducted as an observational, quantitative, retrospective and longitudinal study. It took place in Catalonia, Spain from 2009 to 2019 and was based on the information concerning air pollutants that had been collected by 94 monitoring stations located throughout Catalonia (n = 10,081 records in total). The map in Fig. 1 shows the location of the monitoring stations with respect to the areas in the Catalan health-zone system. We obtained information on the hourly levels of air pollution from the Catalan Network for Pollution Control and Prevention (XVPCA) (open data) (Departament de Territori i Sostenibilitat, 2021). Less than a third of the health zones into which Catalonia is divided, have at least one air pollution monitoring station (105 from a total of 376). One health zone has five monitoring stations, six have three, 22 have two, and the remaining 76 have only one station. The pollutants included in the analysis are coarse particles (PM_{10}), nitrogen dioxide (NO_2), ozone (O_3), sulphur dioxide (SO_2) - all of which are expressed as $\mu g/m^3$ -, and carbon monoxide (CO) - expressed in mg/m^3. From the hourly data we obtained the daily data and from these we calculate the monthly data (in both cases, using an arithmetic mean).

2.2. Compositional analysis

2.2.1. Compositional data analysis of air pollutant concentrations

We model air pollutant concentrations statistically assuming them to be compositional data - CoDa (AL-Dhurafi et al., 2018; Gibergans-Báguena et al., 2020; Jarauta-Bragulat et al., 2016; Sánchez-Balseca and Pérez-Foguet, 2019, 2020). CoDa is the standard approach to analysing the concentrations of parts of a whole (Aitchison, 1986; Boogart et al., 2013; Filzmoser et al., 2018; Greenacre, 2018; Pawlowsky-Glahn et al., 2015a), which is the case not only for air pollution but also for soil, water, and smoke compositions (Bondu et al., 2020; Hron et al., 2021; Karakan et al., 2021; Strbova et al., 2021; Weise et al., 2020).

The usefulness of the CoDa approach lies not only the treatment of parts of a whole, but also in the analysis of data for which the relative importance of magnitudes is of interest, be they parts of a whole or not (Egozcue and Pawlowsky-Glahn, 2019). Since not all pollutants are equally harmful for health and the relative importance of pollutants differs for each health outcome, in the context of pollution studies CoDa makes it possible to identify patterns with differing relative importance of pollutants, and different health risks (Tepanosyan et al., 2021).

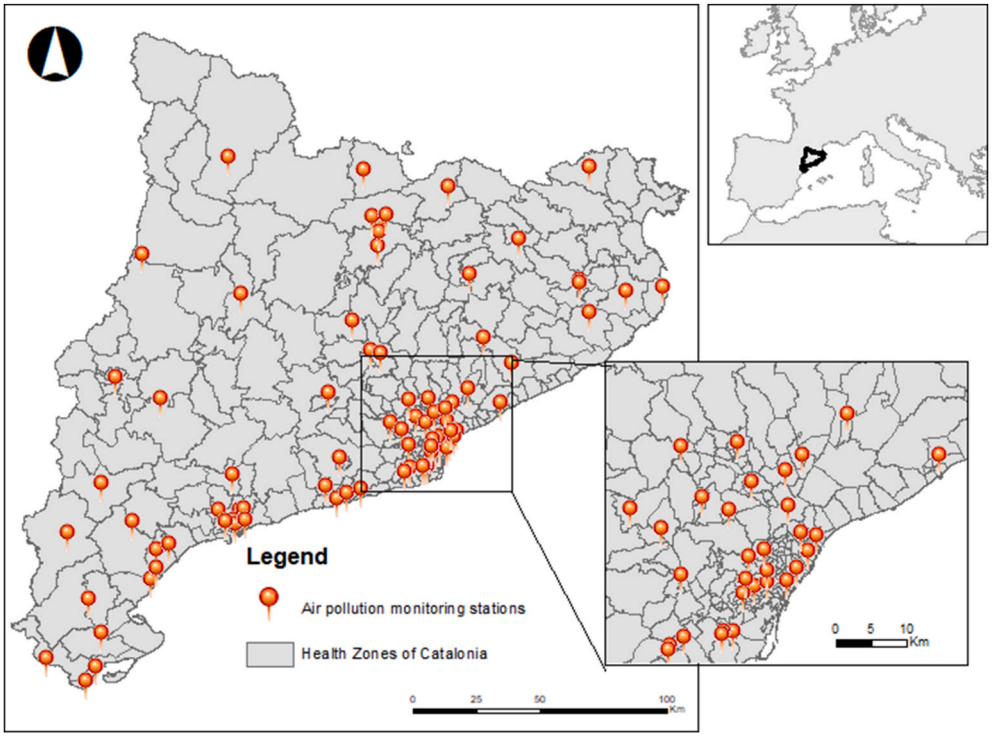


Fig. 1. Distribution of the air pollution monitoring stations with respect to the health zones of Catalonia (Spain).

Opting for the CoDa approach, therefore, not only depends on the nature of the data, but also on the research questions being posed and the way in which the researcher wishes the analysis results to be presented, i.e., as absolute or relative. The term *compositional analysis* has even been coined as an alternative to the term CoDa to stress this fact (Barceló-Vidal and Martín-Fernández, 2016).

The five parts in the composition are the five air pollutants $x_1 = PM_{10}$ ($\mu g/m^3$), $x_2 = NO_2$ ($\mu g/m^3$), $x_3 = O_3$ ($\mu g/m^3$), $x_4 = CO$ (mg/m^3), and $x_5 = SO_2$ ($\mu g/m^3$). In CoDa, the measurement units do not necessarily have to be the same for all parts.

To perform the analysis, we use the program CoDaPack (Comas-Cufí and Thió-Henestrosa, 2011).

The steps in the analysis using CoDaPack are as follows:

1. Imputation of missing values with the log-ratio EM algorithm using the robust option. This is an adaptation of the common algorithm for the replacement of zeros below a detection limit (Palarea-Albaladejo and Martín-Fernández, 2015) implemented by removing the constraint that the imputed values should be below the detection limit. In our case, prior to this step, cases with two or fewer observed contaminants were dropped from the analysis. This resulted in 5498 useable cases from 64 monitoring stations, which were then submitted to the imputation procedure. There were 2460 missing values in PM_{10}, 1044 in O_3, 2055 in CO, and 674 in SO_2. NO_2 had complete information.
2. Outlier detection by means of Mahalanobis distances (Filzmoser et al., 2005), with the cut-off criterion adjusted for sample size, computed as $0.95^{(1/n)}$ (Coenders and Saez, 2000). Here, 88 outliers were identified, resulting in a final sample size of 5410 cases.
3. Exploratory analysis of the composition by means of CoDa biplots (Aitchison and Greenacre, 2002) and the variation matrix (Aitchison, 1986). The variation matrix is a substitute for the correlation matrix in CoDa. As a standard biplot, the CoDa biplot is based on a principal component analysis and presents variables (in this case pollutants) as rays and observations (by place and time) as points and makes it possible to visualise the relationships among variables.
4. Computation of interpretable balance coordinates according to the biplot. The biplot dimensions tend to be difficult to interpret because all dimensions are related to all parts (Martín-Fernández et al., 2017). Balance coordinates obtained from a sequential binary partition of parts (Egozcue and Pawlowsky-Glahn, 2005) represent trade-offs between subsets of parts defined by the user and are much more readily interpretable. More particularly, balance coordinates are scaled log-ratios of the geometric means of the concentrations of two subsets of pollutants and indicate the relative importance of the pollutants in the numerator as compared to those in the denominator of the log-ratio. The easiest-to-interpret balance coordinates are those with one pollutant in the numerator and the rest in the denominator, which can be understood as the relative importance of the pollutant in the numerator within the composition (Fišerová and Hron, 2011; Filzmoser et al., 2018), and those with one air pollutant in the numerator and one in the denominator, which can be understood as the trade-offs between two pollutants (Greenacre, 2018, 2019; Hron et al., 2021). Besides their interpretation in themselves as trade-offs between pollutants, these balance coordinates play the role of variables in any further statistical analysis (Pawlowsky-Glahn et al., 2015a). Balance coordinates are thus the main outcome of the CoDa methodology in the sense that the spatio-temporal model uses them as the outcomes to be predicted.

2.2.2. Introduction of total air pollution levels in CoDa

We wanted to consider not only the relative importance of air pollutants (as in standard CoDa) but overall air pollution, which is also an essential variable in health studies. In the following lines we suggest an analogous to the air quality indices which is coherent with the CoDa methodology.

Implicitly, in an air pollutant concentration composition with D air pollutants, with D equal to 5, there is a residual part x_{D+1} which corresponds to "clean air". Composition x_1 to x_{D+1} may have a fixed sum, e.g., 1,000,000 if all parts are in ppm. However, subcomposition $x_1, x_2 \ldots x_D$ can never have a fixed sum.

One possible approach for the purpose of taking overall air pollution into account is to use the whole composition x_1 to x_{D+1} (Sánchez-Balseca et al., 2020). Under this approach, the balance coordinates involving only the subcomposition x_1 to x_D are used to extract the relative importance of air pollutants to one another as in standard CoDa. An added balance coordinate comparing the geometric average of x_1 to x_D with clean air (x_{D+1}) speaks of overall air pollution levels:

$$\sqrt{\frac{D}{D+1}}\ln\left(\frac{(x_1 \ldots x_D)^{1/D}}{x_{D+1}}\right) = \sqrt{\frac{D}{D+1}}\left(\ln(x_1 \ldots x_D)^{1/D} - \ln(x_{D+1})\right) \quad (1)$$

An alternative approach is to use a T-space (Pawlowsky-Glahn et al., 2015b; Ferrer-Rosell et al., 2016; Coenders et al., 2017) on subcomposition $x_1 \ldots x_D$. A T-space is also known as CoDa with a total, and simply adds some form of total to the subcomposition, so that $x_1 \ldots x_D$ are used to extract the relative importance of air pollutants to one another as before, while the total T speaks of overall air pollution levels. The total is more appropriately defined from the geometric average of air pollutant concentrations than from their sum (Coenders et al., 2017) with a scaling constant to take the number of parts into account:

$$\sqrt{D}\ln\left((x_1 \ldots x_D)^{1/D}\right) \quad (2)$$

It can be argued that the whole air is what is compositional data, thus favouring the whole-composition approach including x_{D+1} (Sánchez-Balseca et al., 2020). However, this loses relevance if x_{D+1} is very large compared to $x_1 \ldots x_D$. Then, ln (x_{D+1}) is nearly constant (Jarauta-Bragulat et al., 2016), and Equations (1) and (2) are close to being linearly related (Martín-Fernández et al., 2020), thus having nearly identical relationships to any external variable. In this context, the advantage of the T-space approach is in avoiding the need to measure the clean air residual part, which may be challenging, especially if air pollutants combine gases, aerosols and particles or are measured in different units. In what follows, we use the T-space approach with total air pollution computed from a geometric mean, as in Equation (2).

The use of the geometric mean of air pollutant concentrations as an overall air pollution measure was first suggested by Jarauta-Bragulat et al. (2016), but these authors did not take balance coordinates into account together with the total, which is our novel contribution. Once in a log-scale, as in Equation (2), changing the units of measurement of the concentration of any air pollutant only results in adding a constant, thus leaving the relationship of the total with any external variable invariant. It is thus not serious if, for instance, particles are measured with different units than gases. This property is not shared with sums or weighted sums of air pollutant concentrations.

Balance coordinates and the total in Equation (2) are used as variables in the spatio-temporal model.

2.3. Bayesian analysis

2.3.1. Specification of the model

We specify a hierarchical spatio-temporal model with the following measurement equation:

$$y(s_i, t) = \mu(s_i, t) + \delta(s_i, t) \quad (3)$$

where $y(.,.)$, is the realization of the spatio-temporal process; $\mu(.,.)$ denotes the large-scale component, depending on the covariates; and $\delta(.,.)$ is a spatio-temporal process, independent in time Gaussian field (GF) with zero mean and a Matérn covariance function (Saez and Barceló, 2021).

Due to its computational problems, we choose to represent the GF as

a Gaussian Markov Random Field (GMRF) (Rue et al., 2009). We link the GF and GMRF through the Stochastic Partial Differential Equations (SPDE) approach (Lindgren et al., 2011). Further details can be found in Saez and Barceló (2021).

We specify the large-scale component, $\mu(.,.)$, as a generalized linear mixed model (GLMM) with response from the Gaussian family:

$$\mu_{i,t} = \beta_0 + \beta_1 altitude_i + \beta_2 area_i + sd_y_{i,year} + \eta_i + \tau_{month} \quad (4)$$

where i denotes the air pollution monitoring station where the pollutant was observed (i = 1,2, …,64); t is the time unit (month in our case); $\mu_{i,t} = E(y_{it})$, y_{it} denotes a balance coordinate or the total in Equation (2); $sd_y_{i,year_{..}}$, η_i and τ_{month} denote random effects.

In all models, we include the $sd_y_{i,year}$ structured random effects, indexed on a standard deviation of the variable that is being predicted in the health zone where the monitoring station is located during a particular year (2009–2019). We choose a random walk of order 1 (rw1) as the structure of the random effect. In the integrated nested Laplace approximations (INLA) approach (Rue et al., 2009, 2017, 2017), the random walk of order 1 for the Gaussian vector z is constructed assuming independent increments (R INLA project, 2021a):

$$\Delta z_i = z_i - z_{i-1} \sim N(0, \sigma_z^2) \quad (5)$$

Following the INLA approach, when, as in our case, the random effects are indexed on a continuous variable, they can be used as smoothers to model non-linear dependency on covariates in the linear predictor.

η_i denotes a random effect indexed on the air pollution monitoring station. This random effect is unstructured (independent and identically distributed) and captures individual heterogeneity; that is to say, unobserved confounders specific to the station and invariant in time.

We also include the τ_{month}, structured random effects indexed on time in order to control the temporal dependency associated to possible seasonal effects throughout the year. In this case, a model for seasonal variation with periodicity m (12 for long-term exposure, seven for short-term exposure), for the random vector (z_1, z_2, …,z_n) ($n > m$) is obtained assuming that the sums are independent Gaussian with a precision τ. The probability density for z is derived from the n-m+1 increments (R INLA project, 2021b):

$$\tau^{\frac{n-m+1}{2}} e^{-\frac{\tau}{2}\sum (z_i + z_{i+1} + \ldots + z_{i+m-1})^2} \quad (6)$$

2.3.2. Inference

Inferences for GMRFs were made following a Bayesian perspective using the INLA approach (Rue et al., 2009, 2017, 2017). We started from the SPDE representation (Lindgren and Rue, 2011) Then, instead of projecting the subsequent mean of the random field onto mesh nodes to target locations where we do not have observed data, we performed a spatial prediction of the random field jointly with the parameter estimation process (Krainski et al., 2020; Saez and Barceló, 2021).

We used priors that penalize complexity (called PC priors). These priors are robust in the sense that they do not have an impact on the results and, furthermore, they have an epidemiological interpretation (Simpson et al., 2017).

All analyses were carried out using the free software R (version 4.0.3), through the INLA package.

2.3.3. Out of sample predictive performance

The INLA method provides the predictions for the balance coordinates and total re-expressed as compositional data. Subsequently, two out-of-sample predictions are made, one for 2019 considering all pollutant monitoring stations and the other for the 2009–2019 period, randomly leaving out 30% of the stations. The measures of predictive accuracy are:

Root mean square error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_i \sum_t (y(s_i,t) - \hat{y}(s_i,t))^2} \quad (7)$$

where $y(s_i,t)$ denotes the observed balance coordinates and total; $\hat{y}(s_i,t)$ the predicted balance coordinates and total; N the number of predicted (observed) balance coordinates and total; i denotes the air pollution monitoring station where the pollutant was observed (i = 1,2, …,64) and t is the time unit (month in our case).

The RMSE can be compared to the standard deviation of the original data.

Product-moment correlation between the predictions and the original data:

$$r = \frac{\sum_i \sum_t \left(y(s_i,t) - \overline{y(s_i,t)}\right)\left(\hat{y}(s_i,t) - \overline{\hat{y}(s_i,t)}\right)}{\left(\sum_i \sum_t \left(y(s_i,t) - \overline{y(s_i,t)}\right)^2 \sum_i \sum_t \left(\hat{y}(s_i,t) - \overline{\hat{y}(s_i,t)}\right)^2\right)^{\frac{1}{2}}} \quad (8)$$

where $\overline{y(s_i,t)}$ denotes the mean of the observed balance coordinates and total and $\overline{\hat{y}(s_i,t)}$ the mean of the predicted balance coordinates and total.

3. Results

3.1. Compositional data results

The elements in the variation matrix (Table 1) are the variances of the log-ratios between pairs of air pollutants. Low values indicate pairs of air pollutants which move proportionally (Pawlowsky-Glahn et al., 2015a).

We can see the highest log-ratio variance is between O_3 and NO_2. This means that monitoring stations and periods with high levels of O_3 tend to have low levels of NO_2, and vice-versa. It is well known that NO_2 concentrations increase in urban areas and during the summer months while the opposite holds for O_3. The lowest log-ratio variance is between CO and PM_{10}, and for this reason we could say that these two air pollutants move proportionally: monitoring stations and periods with high levels of CO tend to have also high levels of PM_{10}.

In the case of SO_2 and NO_2 or SO_2 and O_3, there are also high log-ratio variance values, though not as much as between O_3 and NO_2. Other cases with low log-ratio variances are NO_2 and PM_{10} or O_3 and PM_{10}.

The first two dimensions of the principal component analysis that are represented in the covariance CoDa biplot explain 86.3% of the variance. In the biplot in Fig. 2, distances between rays corresponding to air pollutants are approximately proportional to the square root of log-ratio variance. The high log-ratio variance between NO_2 and O_3 can be clearly observed as they are completely opposite in their respective directions, whereas PM_{10} and CO are closer together. The horizontal axis opposes O_3 with NO_2, while the vertical axis opposes $S0_2$ with all the remaining air pollutants. The presence of SO_2 increases in industrial areas while most of the other air pollutants are related to the road traffic.

Table 2 shows the sign matrix in the sequential binary partition leading to the balance coordinates. Values −1 or 1 indicate whether the air pollutants are in the denominator or the numerator of the balance coordinates. Zeros indicate pollutants that are neither in the numerator nor in the denominator. According to the biplot, the first partition in the

Table 1
Variation matrix.

Air pollutants	PM_{10}	NO_2	O_3	CO	SO_2
x_1: PM_{10}		0.45	0.35	0.12	0.40
x_2: NO_2	0.45		1.41	0.50	0.75
x_3: O_3	0.35	1.41		0.66	0.73
x_4: CO	0.12	0.50	0.66		0.45
x_5: SO_2	0.40	0.75	0.73	0.45	

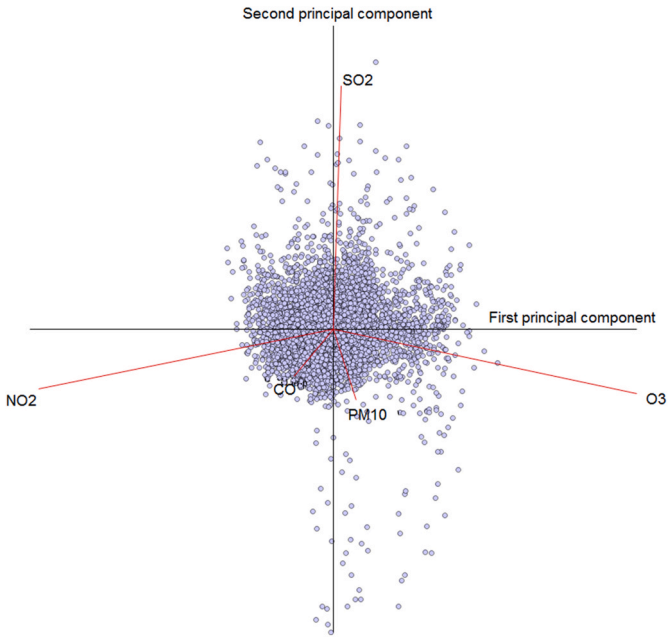


Fig. 2. Compositional covariance biplot.

Table 2
Sign matrix of the binary partition.

x_1: PM_{10}	x_2: NO_2	x_3: O_3	x_4: CO	x_5: SO_2
−1	−1	−1	−1	1
−1	1	1	−1	0
0	1	−1	0	0
1	0	0	−1	0

sign matrix (Table 2) places SO_2 in the numerator and the remaining air pollutants in the denominator and is a simplification of the second dimension in the biplot. The third partition places NO_2 in the numerator and O_3 in the denominator and parallels the first dimension.

The corresponding balance coordinates are the log-ratios of the geometric means of the involved numerator and denominator parts with a scaling constant taking into account the number of parts involved (Egozcue and Pawlowsky-Glahn, 2005) and are:

$$x_1^* = \sqrt{\frac{4}{5}}\ln\left(\frac{x_5}{(x_1x_2x_3x_4)^{1/4}}\right)$$

$$x_2^* = \sqrt{\frac{4}{4}}\ln\left(\frac{(x_2x_3)^{1/2}}{(x_1x_4)^{1/2}}\right)$$

$$x_3^* = \sqrt{\frac{1}{2}}\ln\left(\frac{x_2}{x_3}\right)$$

$$x_4^* = \sqrt{\frac{1}{2}}\ln\left(\frac{x_1}{x_4}\right) \quad (9)$$

where x_1 stands for PM_{10}, x_2 NO_2, x_3 O_3, x_4 CO and x_5 SO_2.

The two main balance coordinates (x^*_1 and x^*_3) explain 85.6% of log-ratio variance, which compares quite well with the optimal 86.3% explained by the first two principal components in the biplot. They represent the main sources of variability in pollution concentrations, together with the total in Equation (2). High values of x^*_1 indicate a high relative importance of SO_2 with respect to NO_2, O_3, PM_{10} and CO. High values of x^*_3 indicate high importance of NO_2 as compared to O_3.

The four balance coordinates x^*_1 to x^*_4 in Equation (9) and the total in Equation (2) are the raw data in the Bayesian hierarchical spatio-temporal model used to predict air pollution in locations or time periods without pollution monitoring sites. Table 3 shows the descriptive statistical analysis and we can see that the highest variances are found in the balance coordinates x^*_1, x^*_3 and the total (T).

Fig. 3 shows the boxplots of the relative importance of pollutants as expressed by x^*_1 and x^*_3 and total pollution (T), for the Barcelona County (Barcelonès) and other Catalan counties (referred to as Comarcas). If we look at the Barcelonès, apart from presenting a higher pollution globally, it highlights a lower relative concentration of SO_2 as compared to the remaining pollutants, and also an exchange between O_3 and NO_2.

In Fig. 4, the total pollution is higher in the winter and during the rest of the quarters it is lower. On the other hand, if we look at x^*_3, the relationship between NO_2 and O_3 changes during the autumn-winter months in favour of NO_2, and during the spring-summer months in favour of O_3.

Fig. 5 shows that the total pollution has substantially decreased between 2009 and 2019. If we look at the x^*_3 balance, in 2009 the ratio of NO_2 to O_3 was higher than in 2019. In addition, in the case of the x^*_1 balance, the relationship between SO_2 and the rest of the pollutants has stayed approximately constant over the two years plotted.

3.2. Predictive accuracy

Table 4 shows the root mean square error (RMSE, Equation (7)), the original standard deviation (SD) and the correlation (Equation (8)) of the respective balance coordinates, and the total when leaving out the final year (2019). x^*_3 and total air pollution have substantially lower RMSE than SD values and a high correlation coefficient, which argues for a very high predictive accuracy for out-of-sample future time periods. Predictive accuracy is very low for x^*_1. The remaining balance coordinates have a moderate predictive accuracy.

In the following assessment, as mentioned above, the predictions are made throughout the 2009–2019 period but 30% of the pollutant monitoring stations have been randomly left out (Table 5). As in the previous case, what is interpreted is the reduction between the RMSE and the SD, as well as the correlation between the actual and predicted values for the respective balance coordinates and the total. The total now has the poorest predictive accuracy, while the balance coordinates have improved theirs, especially x^*_1. As an example, Table 6 shows six monitoring stations chosen at random with their respective raw data and out-of-sample predictions for 2018.

4. Discussion

Recently, the distribution in space and time of pollutants in Catalonia has attracted great interest and a variety of methods beyond the CoDa methodology and hierarchical Bayesian spatio - temporal models have been employed for a variety of scientific and policy-making purposes. To highlight but a few, this includes, for instance, the effects the COVID-19 pandemic lockdown had on pollution (Baldasano, 2020; Tobías et al., 2020), pollution in rural areas (Jaén et al., 2021), citizen science campaigns to measure pollution (Perelló et al., 2021), or the effects of public transport strikes (González et al., 2021).

From the approach we suggest in this article, the compositional analysis has made it possible to identify the main sources of log-ratio

Table 3
Descriptive statistics of balance coordinates and the total air pollution indicator.

	Mean	Variance	Min	Q1	Q2	Q3	Max
x^*_1	−1.23	0.29	−5.40	−1.52	−1.25	−0.98	2.48
x^*_2	2.50	0.11	0.75	2.38	2.54	2.68	4.07
x^*_3	−0.64	0.70	−4.17	−1.09	−0.55	−0.08	1.66
x^*_4	2.98	0.06	1.72	2.88	3.00	3.11	4.18
T	4.23	0.57	−0.004	3.94	4.36	4.72	6.37

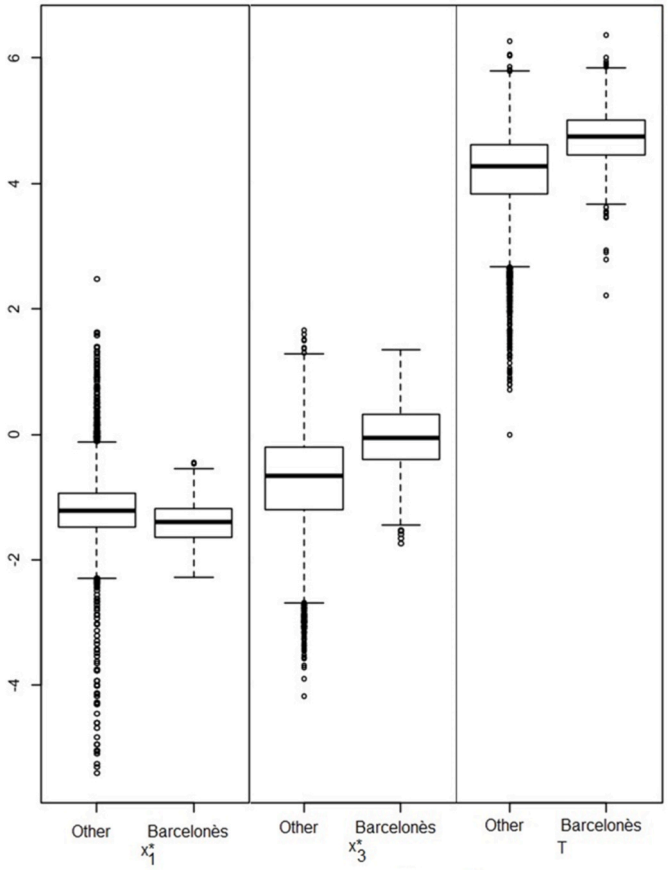


Fig. 3. Boxplots of the balance coordinates x^*_1, x^*_3 and the total by location (Barcelonès versus other areas) during the period 2009–2019.

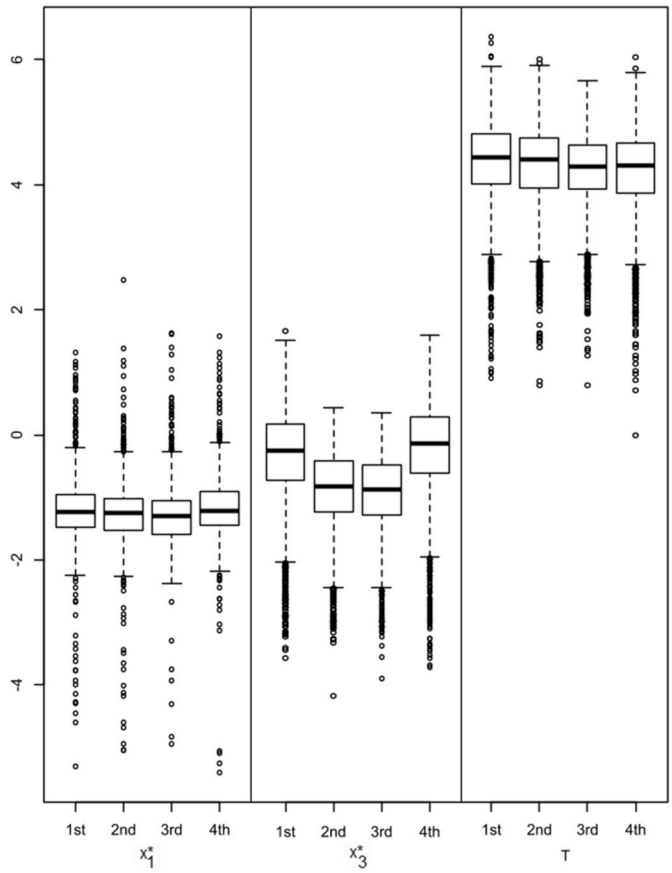


Fig. 4. Boxplots of the balance coordinates x^*_1, x^*_3 and the total by quarters during the period 2009–2019, Note: First quarter: January to March; Second quarter: April to June; Third quarter: July to September; Fourth quarter: October to December.

variance. Among all the pairs of air pollutants studied, the pair with the most variability is that formed by O_3 and NO_2. This variance explains that when high concentrations of O_3 are given, low concentrations of NO_2 are obtained and vice-versa.

The log-ratio of SO_2 over the remaining pollutants comes second in explained variance. This fact could be related to the strong presence of SO_2 in industrial and non-urban areas. On the other hand, it shows that other pollutants are higher in areas where there is a high amount of road traffic.

The CoDa methodology through log-ratios, with the corresponding balance coordinates, makes it possible to highlight trade-offs between pollutants or groups of pollutants and is used in other studies such as those carried out by AL-Dhurafi et al. (2018) and Sánchez -Balseca and Pérez-Foguet (2019, 2020). This focus on the trade-offs is compatible to an overall air quality index by means of which we have termed total. Gibergans-Báguena et al. (2020) and Jarauta et al. (2016) consider only the total. AL-Dhurafi et al. (2018); Sánchez-Balseca and Pérez-Foguet (2019); Gibergans-Báguena et al. (2020) and Jarauta et al. (2016) use time series models without spatial dimension. Thus, our article is most similar to Sánchez-Balseca and Pérez-Foguet (2020). The main differences are the use of the INLA approach instead of Markov Chain Monte Carlo (MCMC) and expressing total pollution with Equation (2) instead of (1).

Regarding the Bayesian component, our innovations in this paper lie in, on the one hand, that the spatio-temporal approach we use requires data from fewer monitoring stations but yields a precision comparable to other approaches (Saez and Barceló, 2021). On the other hand, in the estimation we use the INLA approximation. Using MCMC implies a high computational model complexity that entails high computational costs. In addition, on some occasions this complexity prevents practical applications particularly when, as in our case, working with a spatio-temporal design with variability in both dimensions. In fact, the INLA approach is much more computationally effective than MCMC, producing accurate approximations to subsequent distributions, even for very complex models (Lindgren and Rue, 2015).

The Bayesian analysis, using the INLA method to predict the pollutant balance coordinates and the total temporally during 2019 confirms, in the present study, that the best correlation between the predicted and actual values occurs in the x^*_3 balance, which is that of NO_2 with respect to O_3, and the worst in the x^*_1 balance that takes into account the relationship between SO_2 with respect to PM_{10}, NO_2, O_3 and CO. Sánchez-Balseca et al. (2020) also obtain poor accuracy in relation to SO_2.

The INLA method has also enabled us to make spatial predictions during the period 2009–2019 by excluding 30% of the monitoring stations at random. As in the time-level analysis mentioned in the previous paragraph, the best correlation is in the x^*_3 balance of NO_2 with respect to O_3. This is fortunate as the x^*_3 balance alone explains 60.5% of the compositional variance. The accuracy for x^*_1 is adequate, but the accuracy for total pollution is extremely poor. As Sánchez-Balseca et al. (2020) do not make spatial prediction, we cannot compare our results to theirs.

Very recently, Saez and Barceló (2021) presented a hierarchical Bayesian spatio-temporal model to perform spatial predictions of air pollution levels. They used the SPDE representation of the INLA approximation to spatially predict in the territory of Catalonia both long and short-term exposure to four pollutants: PM_{10}, NO_2, O_3 and $PM_{2.5}$. In what is comparable (the balance coordinate between NO_2 and O_3), the predictive performance of our study is about the same as in Saez and

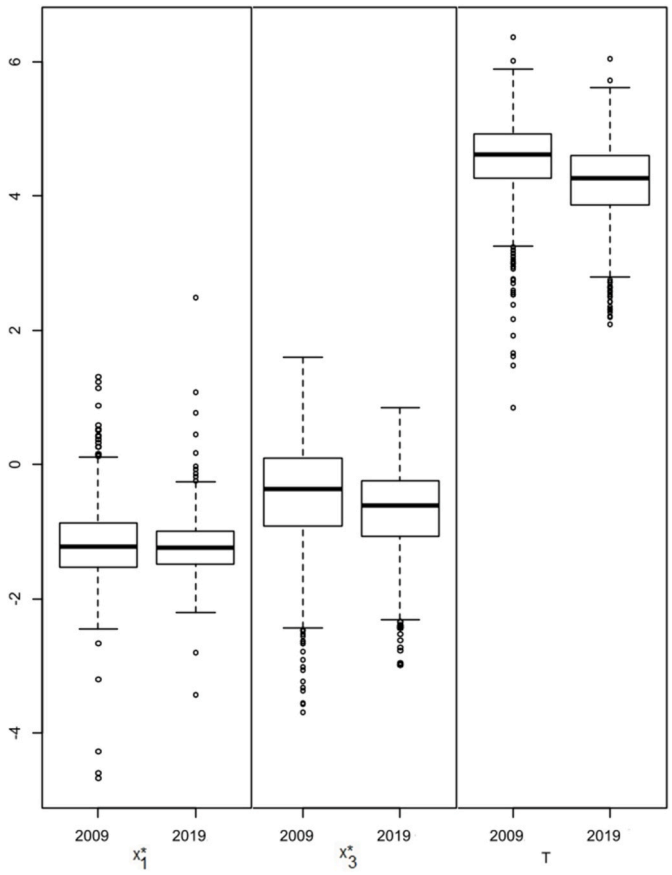


Fig. 5. Boxplots of the balance coordinates x^*_1, x^*_3 and the total by 2009 and 2019.

Table 4
Accuracy of temporal predictions for 2019.

	RMSE[a]	SD [b]	Correlation[c]
x^a_1	0.39	0.48	0.12
x^a_2	0.23	0.39	0.45
x^a_3	0.38	0.71	0.82
x^a_4	0.15	0.27	0.49
T	0.32	0.65	0.83

[a] Root mean squared prediction error.
[b] Standard deviation of the raw data.
[c] Correlation between the predictions and the raw data.

Table 5
Accuracy of spatial predictions for the period 2009–2019 for 30% of stations omitted.

	RMSE[a]	SD[b]	Correlation[c]
x^a_1	0.40	0.54	0.67
x^a_2	0.25	0.33	0.63
x^a_3	0.44	0.84	0.86
x^a_4	0.19	0.24	0.62
T	0.82	0.76	−0.28

[a] Root mean squared prediction error.
[b] Standard deviation of the raw data.
[c] Correlation between the predictions and the raw data.

Barceló (2021). Having said that, the aim of the CoDa approach is not to achieve higher precision and it is very important to note that this method allows variables to be presented differently by separating overall pollution from trade-offs between pollutants.

The pollutant SO_2 does not yield reliable or accurate predictions for new time periods but is well predicted in sites without a monitoring station. Thus, the SO_2 concentration has hard-to-predict time variability. The main source of SO_2 is from industry and their locations have been taken away from urban and peri-urban areas which, in any case, do not have nearby pollution monitoring stations. In addition, the fight against air pollution has meant filters and other decontaminating measures are being employed. In recent years there has been a sharp decrease in SO_2 emissions produced by combustion in energy production industries due to various factors; for instance, the 2007 Plan for the Reduction of Emissions from Large Combustion Facilities (Government of Spain), which forced the introduction of desulfurization technologies (Ministry for the Ecological Transition and the Demographic Challenge, 2021). On the other hand, the temporal distribution of SO_2 levels is usually determined by the periods in which the emission industry that affects the station is in operation, which would imply great temporal variability (Ministry for Ecological Transition and Demographic Challenge, 2021).

The total is well predicted in the case of new time periods but is difficult to make spatial predictions if there is no monitoring station nearby because of the high spatial variability of total pollution. On the contrary, the relationship between air pollutants contained in the balance coordinates is easy to predict in new monitoring sites because it depends on climatic and seasonal variations.

The use of the CoDa method to analyse air pollution allows for a clearer understanding of the data because it shows trade-offs between air pollutants besides overall pollution. The balance coordinates and the total can be used as variables in the spatio-temporal analysis with the application of the INLA method to make effective predictions of air pollution.

From among the Bayesian methods, we chose the INLA approach because using Monte Carlo Markov Chain (MCMC) methods implies a high computational model complexity that, in some cases, prevents the practical application of these methods or restricts the researcher to simpler model specifications. In fact, compared to MCMC, INLA allows spatial predictions of air pollution levels to be made in a more effective way and with considerably less computational cost.

The limitations of the study are two-fold: the number of missing values at the various monitoring stations and the non-homogeneous distribution of the stations themselves as they are concentrated in certain geographical areas linked to higher population density or to more intense industrial activity (Fig. 1). Taken together, both limitations leave sizeable gaps in the monitoring-station map, and contribute to the poor spatial prediction of total pollution levels.

Funding

This work was partially financed by the SUPERA COVID19 Fund, from SAUN: Santander Universidades, CRUE and CSIC; by the COVID-19 Competitive Grant Program from Pfizer Global Medical Grants; by the Spanish Ministry of Science, Innovation and Universities/FEDER (grant number RTI 2018–095518–B–C21); and the Government of Catalonia (grant number 2017SGR656). The funding sources did not participate in the design or conduct of the study, the collection, management, analysis, or interpretation of the data, nor the preparation, review, or approval of the manuscript.

Ethics

Not applicable.

Software and data availability

We used open data with free access from these sources:

Departament de Territori i Sostenibilitat, Generalitat de Catalunya [Available at: https://analisi.transparenciacatalunya.cat/en/Medi-Ambient/Qualitat-de-l-aire-als-punts-de-mesurament-autom-t/tasf-th

Table 6
Examples of raw data and spatial predictions for six monitoring stations (2018).

Monitoring Station	x_1^*	$\widehat{x}_1^*$	x_2^*	$\widehat{x}_2^*$	x_3^*	$\widehat{x}_3^*$	x_4^*	$\widehat{x}_4^*$	T	$\widehat{T}$
Sant Domènec-Itàlia, Amposta	−1.198	−1.199	2.654	2.655	−1.074	−1.075	3.112	3.111	3.849	3.850
Gràcia-Sant Gervasi, Barcelona	−1.575	−1.576	2.617	2.620	0.105	0.106	2.865	2.860	4.692	4.693
CEIP Mare de Déu de Talló, La Cerdanya	−1.176	−1.187	2.768	2.764	−1.130	−1.124	3.188	3.191	3.138	3.156
Laboratori d'Aigües, Mataró	−0.781	−0.782	2.816	2.817	−0.648	−0.650	2.906	2.910	4.489	4.490
Zona Esportiva, Tona	−1.282	−1283	2.474	2.475	−2.353	−2.354	3.130	3.131	2.350	2.351
RENFE, Vila-seca	−0.993	−0.994	2.788	2.789	−0.791	−0.792	3.158	3.159	4.043	4.044

gu, last accessed on March 14, 2021].

Digitized cartography of the Catalan health zones. Departament de Salut. Cartography [Available at: https://salutweb.gencat.cat/ca/el_departament/estadistiques_sanitaries/cartografia/, accessed on March 14, 2021].

CoDaPack can be downloaded at http://ima.udg.edu/codapack/

R code will be available at www.researchprojects.es.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We appreciate the comments made by the three anonymous reviewers of a previous version of this work who, without doubt, helped us to improve our work. The usual disclaimer applies. We also thank Diego Varga (GRECS, University of Girona and CIBERESP) for his help in the construction of the maps. This study was carried out within the 'Cohort-Real World Data' subprogram of CIBER of Epidemiology and Public Health (CIBERESP).

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman and Hall, London. https://www.jstor.org/stable/2345821.

Aitchison, J., Greenacre, M., 2002. Biplots of compositional data. J. Royal Stat. Soc. Appl. Stat. Ser. 51 (4), 375–392. https://doi.org/10.1111/1467-9876.00275.

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. https://doi.org/10.1007/s00477-018-1542-0.

Baldasano, J.M., 2020. COVID-19 lockdown effects on air quality by NO2 in the cities of Barcelona and Madrid (Spain). Sci. Total Environ. 741, 140353. https://doi.org/10.1016/J.SCITOTENV.2020.140353.

Barceló-Vidal, C., Martín-Fernández, J.A., 2016. The mathematics of compositional analysis. Austrian J. Stat. 45 (4), 57–71. https://doi.org/10.17713/ajs.v45i4.142.

Blangiardo, M., Finazzi, F., Cameletti, M., 2016. Two-stage Bayesian model to evaluate the effect of air pollution on chronic respiratory diseases using drug prescriptions. Spatial Spatio Temp. Epidemiol. 18, 1–12. https://doi.org/10.1016/j.sste.2016.03.001.

Bondu, R., Cloutier, V., Rosa, E., Roy, M., 2020. An exploratory data analysis approach for assessing the sources and distribution of naturally occurring contaminants (F, Ba, Mn, As) in groundwater from southern Quebec (Canada). Appl. Geochem. 114, 104500. https://doi.org/10.1016/j.apgeochem.2019.104500.

Boogaart, K.G. Van den, Tolosana-Delgado, R., 2013. Analyzing Compositional Data with R. Springer, Berlin, pp. 73–93. https://doi.org/10.1007/978-3-642-36809-7_4.

Cameletti, M., Ignaccolo, R., Bande, S., 2011. Comparing spatio-temporal models for particulate matter in Piemonte. Environmetrics 22 (8), 985–996. https://doi.org/10.1002/env.1139.

Cameletti, M., Lindgren, F., Simpson, D., Rue, H., 2013. Spatio-temporal modelling of particular matter concentration through the SPDE approach. AStA Adv. Stat. Anal. 97, 109–131. https://doi.org/10.1007/s10182-012-0196-3.

Coenders, G., Martín-Fernández, J.A., Ferrer-Rosell, B., 2017. When relative and absolute information matter: compositional predictor with a total in generalized linear models. Stat. Model. Int. J. 17 (6), 494–512. https://doi.org/10.1177/1471082X17710398.

Coenders, G., Saez, M., 2000. Collinearity, heteroscedasticity and outlier diagnostics in regression. Do they always offer what they claim? In: Ferligoj, A., Mrvar, A. (Eds.), New Approaches in Applied Statistics, Metodološki Zvezki. 16. FDV, Ljubljana, pp. 79–94 (SI). http://dk.fdv.uni-lj.si/metodoloskizvezki/Pdfs/Mz16CoendersSaez.pdf.

Comas-Cufí, M., Thió-Henestrosa, S., 2011. CoDaPack 2.0: a stand-alone, multi-platform compositional software. In: Egozcue, J.J., Tolosana-Delgado, R., Ortego, M.I. (Eds.), CoDaWork'11: 4th International Workshop on Compositional Data Analysis. Sant Feliu de Guíxols. https://www.researchgate.net/publication/266172921.

Departament de Territori i Sostenibilitat, 2021. Generalitat de Catalunya [Available at: https://analisi.transparenciacatalunya.cat/en/Medi-Ambient/Qualitat-de-l-aire-als-punts-de-mesurament-autom-t/tasf-thgu. (Accessed 11 May 2021).

Egozcue, J.J., Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. Math. Geol. 37, 795–828. https://doi.org/10.1007/s11004-005-7381-9.

Egozcue, J.J., Pawlowsky-Glahn, V., 2019. Compositional data: the sample space and its structure. Test 28, 599–638. https://doi.org/10.1007/s11749-019-00670-6.

European Environment Information and Observation Network (EIONET), 2020. Air Pollutants [Available at: https://www.eionet.europa.eu/gemet/es/concept/263. (Accessed 30 April 2021).

Ferrer-Rosell, B., Coenders, G., Mateu-Figueras, G., Pawlowsky-Glahn, V., 2016. Understanding low-cost airline users' expenditure patterns and volume. Tourism Econ. 22 (2), 269–291. https://doi.org/10.5367/te.2016.0548.

Filzmoser, P., Garrett, R.G., Reimann, C., 2005. Multivariate outlier detection in exploration geochemistry. Comput. Geosci. 31 (5), 579–587. https://doi.org/10.1016/j.cageo.2004.11.013.

Filzmoser, P., Hron, K., Templ, M., 2018. Applied Compositional Data Analysis with Worked Examples in R. Springer, New York. https://www.springer.com/gp/book/9783319964201.

Fišerová, E., Hron, K., 2011. On interpretation of orthonormal coordinates for compositional data. Math. Geosci. 43 (4), 455–468. https://doi.org/10.1007/s11004-011-9333-x.

Gibergans-Báguena, J., Hervada-Sala, C., Jarauta-Bragulat, E., 2020. The quality of urban air in Barcelona: a new approach applying Compositional Data Analysis Methods. Emerg. Sci. 4 (2) https://doi.org/10.28991/esj-2020-01215.

González, L., Perdiguero, J., Sanz, À., 2021. Impact of public transport strikes on traffic and pollution in the city of Barcelona. Transport. Res. Transport Environ. 98, 102952. https://doi.org/10.1016/J.TRD.2021.102952.

Greenacre, M., 2018. Compositional Data Analysis in Practice. Chapman and Hall/CRC press, New York. https://doi.org/10.1201/9780429455537.

Greenacre, M., 2019. Variable selection in compositional data analysis using pairwise logratios. Math. Geosci. 51 (5), 649–682.

Hron, K., Coenders, G., Filzmoser, P., Palarea-Albaladejo, J., Faměra, M., Matys-Grygar, T., 2021. Analysing pairwise logratios revisited. Math. Geosci. https://doi.org/10.1007/s11004-021-09938-w.

Jaén, C., Villasclaras, P., Fernández, P., Grimalt, J.O., Udina, M., Bedia, C., Drooge, B. L. van, 2021. Source apportionment and toxicity of PM in urban, sub-urban, and rural air quality network stations in Catalonia. Atmosphere 2021 12, 744. https://doi.org/10.3390/ATMOS12060744, 12(6), 744.

Jarauta-Bragulat, E., Hervada-Sala, C., Egozcue, J.J., 2016. Air quality index revisited from a compositional point of view. Math. Geosci. 1;48 (5), 581–593. https://doi.org/10.1007/s11004-015-9599-5.

Jia, H., Kikumoto, H., 2021. Line source estimation of environmental pollutants using super-Gaussian geometry model and Bayesian inference. Environ. Res. 194 https://doi.org/10.1016/j.envres.2020.110706.

Karakan, Ö.C., Martín-Fernández, J.A., Ruppert, L.F., Ricardo, A.O., 2021. Insights on the characteristics and sources of gas from an underground coal mine using compositional data analysis. Int. J. Coal Geol. 241 https://doi.org/10.1016/j.coal.2021.103767.

Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., Rue, H., 2020. Advanced Spatial Modelling with Stochastic Partial Differential Equations Using R and INLA. Chapman and Hall/CRC, London. https://doi.org/10.1201/9780429031892 (chapter 2).5.

Liang, Y.C., Maimury, Y., Chen, A.H., Juarez, J.R., 2020. Machine learning-based prediction of air quality. Appl. Sci. 10 (24), 9151. https://doi.org/10.3390/app10249151.

Lindgren, F.K., Rue, H., Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J. Royal Stat. Soc. Stat. Methodol. Series 73 (4), 423–498 j.1467-9868.2011.00777.x.

Lindgren, F., Rue, H., 2015. Bayesian spatial modelling with R-INLA. J. Stat. Software 63 (19). https://doi.org/10.18637/jss.v063.i19.

Liu, F., Zhang, Z., Chen, H., Nie, S., 2020. Associations of ambient air pollutants with regional pulmonary tuberculosis incidence in the central Chinese province of Hubei: a Bayesian spatial-temporal analysis. Environ. Health 19 (1), 51. https://doi.org/10.1186/s12940-020-00604-y.

Martín-Fernández, J.A., Egozcue, J.J., Olea, R.A., Pawlowsky-Glahn, V., 2020. Units recovery methods in compositional data analysis. Nat. Resour. Res. 24, 1–4. https://doi.org/10.1007/s11053-020-09659-7.

Martín-Fernández, J.A., Pawlowsky-Glahn, V., Egozcue, J.J., Tolosona-Delgado, R., 2017. Advances in principal balances for compositional data. Math. Geosci. 1–26. https://doi.org/10.1007/s11004-017-9712-z.

Ministry for the Ecological Transition and the Demographic Challenge, 2021. Government of Spain. Sulphur dioxide [in Spanish] [Available at: https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/salud/dioxido-azufre.aspx. (Accessed 8 November 2021).

Pan American Health Organization (PAHO), 2017. Air quality. Available at: https://www.paho.org/en/topics/air-quality. (Accessed 12 April 2021).

Palarea-Albaladejo, J., Martín-Fernández, J.A., 2015. zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. Chemometr. Intell. Lab. Syst. 143, 85–96. https://doi.org/10.1016/j.chemolab.2015.02.019.

Pawlowsky-Glahn, V., Egozcue, J.J., Lovell, D., 2015b. Tools for compositional data with a total. Stat. Model. Int. J. 15 (2), 175–190. https://doi.org/10.1177/1471082X14535526.

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R., 2015a. Modelling and Analysis of Compositional Data. Wiley, Chichester. https://doi.org/10.1002/9781119003144.

Perelló, J., Cigarini, A., Vicens, J., Bonhoure, I., Rojas-Rueda, D., Nieuwenhuijsen, M.J., Cirach, M., Daher, C., Targa, J., Ripoll, A., 2021. Large-scale citizen science provides high-resolution nitrogen dioxide values and health impact while enhancing community knowledge and collective action. Sci. Total Environ. 789, 147750. https://doi.org/10.1016/J.SCITOTENV.2021.147750.

R INLA project, 2021a. Random Walk of Order 1 (RW1) [Available at: https://inla.r-inla-download.org/r-inla.org/doc/latent/rw1.pdf. (Accessed 11 May 2021).

R INLA project, 2021b. Model for Seasonal Variation. Available at: https://inla.r-inla-download.org/r-inla.org/doc/latent/seasonal.pdf. (Accessed 11 May 2021).

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). J. Royal Stat. Soc. Stat. Methodol. Series 71, 319–392 j.1467-9868.2008.00700.x.

Rue, H., Riebler, A., Sørbye, H., Illian, J.B., Simpson, D.P., Lindgren, F.K., 2017. Bayesian computing with INLA: a review. Ann. Rev. Stat. Appl. 4 (March), 395–421 annurev-statistics-060116-054045.

Saez, M., Barceló, MA., 2021. Spatial prediction of air pollution levels using a hierarchical Bayesian spatio-temporal model in Catalonia, Spain. medRxiv the preprint server for health sciences. https://doi.org/10.1101/2021.06.06.21258419.

Sánchez-Balseca, J., Pérez-Foguet, A., 2019. Assessing CoDa regression for modelling daily multivariate air pollutants evolution. In: Egozcue, J.J., Graffelman, J., Ortego, M.I. (Eds.), Proceedings of the 8th International Workshop on Compositional Data Analysis (CoDaWork2019). Universitat Politècnica de Catalunya-BarcelonaTECH, Terrassa, pp. 143–150. ISSBN: 9788494724022.

Sánchez-Balseca, J., Pérez-Foguet, A., 2020. Spatio-temporal air pollution modelling using a compositional approach. Heliyon 6 (9), e04794. https://doi.org/10.1016/j.heliyon.2020.e04794.

Simpson, D.P., Rue, H., Martins, T.G., Riebler, A., Sørbye, S.H., 2017. Penalising model component complexity: a principled, practical approach to constructing priors (with discussion). Stat. Sci. 32 (1), 1–46. https://doi.org/10.1214/16-STS576.

Sicard, P., Agathokleous, E., De Marco, A., Paoletti, E., Calatayud, V., 2021. Urban population exposure to air pollution in Europe over the last decades. Environ. Sci. Eur. 33 (1) https://doi.org/10.1186/s12302-020-00450-2.

Strbova, K., Rusickova, J., Raclavska, H., 2019. Application of multivariate statistical analysis using organic compounds: source identification at a local scale (Napajedla, Czechia). J. Environ. Manag. 238, 434–441. https://doi.org/10.1016/j.jenvman.2019.03.035.

Tepanosyan, G., Sahakyan, L., Maghakyan, N., Saghatelyan, A., 2021. Identification of spatial patterns, geochemical associations and assessment of origin-specific health risk of potentially toxic elements in soils of Armavir region, Armenia. Chemosphere 262, 128365. https://doi.org/10.1016/j.chemosphere.2020.128365.

Tobías, A., Carnerero, C., Reche, C., Massagué, J., Via, M., Minguillón, M.C., Alastuey, A., Querol, X., 2020. Changes in air quality during the lockdown in Barcelona (Spain) one month into the SARS-CoV-2 epidemic. Sci. Total Environ. 726, 138540. https://doi.org/10.1016/J.SCITOTENV.2020.138540.

Weise, D.R., Jung, H., Palarea-Albaladejo, J., Cocker, D.R., 2020. Compositional data analysis of smoke emissions from debris piles with low-density polyethylene. J. Air Waste Manag. Assoc. 70 (8), 834–845. https://doi.org/10.1080/10962247.2020.1784309.

WHO-World Health Organization, 2018. Ambient (Outdoor) Air Pollution [Available at: https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health. (Accessed 12 April 2021).