Received: 14 December 2022 | Revised: 29 November 2023 | Accepted: 30 November 2023

DOI: 10.1002/env.2837

RESEARCH ARTICLE

WILEY

# Structural equation models for simultaneous modeling of air pollutants

**Mariaelena Bottazzi Schenone**[1] | **Elena Grimaccia**[2] | **Maurizio Vichi**[1]

[1]Department of Statistical Sciences, Sapienza University, Rome, Italy

[2]Istat, Italian National Institute of Statistics, Rome, Italy

**Correspondence**
Elena Grimaccia, Istat, Italian National Institute of Statistics, via Cesare Balbo 16, Rome 00184, Italy.
Email: elgrimac@istat.it

**Abstract**

This paper provides a new modeling for air pollution, simultaneously taking into account the six main pollutants (PM10 and PM2.5, Sulphate Dioxide, Nitrogen Dioxide, Carbon Monoxide, ground level Ozone concentrations) and their key determinants, employing Structural Equation Models (SEMs). The model is able to estimate the complex links among air pollutants, often neglected in literature, and identifies specific drivers of air pollution. In literature, indexes of air pollution achieved using a fully statistical methodology have not been proposed yet. Indeed, an added value of this proposal is the statistical procedure itself, which can be applied also to obtain indexes modeling different phenomena. In particular, in this study, the new Air Pollution Index (API) is based on a modeling approach that allows to assess, through statistical criteria, the goodness of fit of the SEM in modeling pollutants and the significance of their determinants. The performance of the new index is assessed using air quality data for municipal European areas, which are characterized by different socioeconomic, geographical, and meteorological features. SEMs are estimated and evaluated in terms of best fit and model complexity. The index resulting by the best SEM is compared with the well-established Air Quality Index (AQI). The new API is validated by means of a sensitivity analysis, performed with a simulation study. Finally, to visualize the meaningfulness of the obtained results, a model-based cluster analysis is estimated on the municipal areas. The proposed SEM contributes to a better understanding of the relationships between air pollutants and their determinants, and this knowledge can inform policy decisions aimed at reducing air pollution and improving public health.

**KEYWORDS**

air pollution, hierarchical models, metropolitan areas, model-based multidimensional index, structural equation models

## 1 | INTRODUCTION

Air pollution has become a serious threat to human health (Hoskovec et al., 2022), particularly in urban areas and cities (Antanasijević et al., 2018; Cromar et al., 2020; Dominici et al., 2000). The World Health Organization (WHO), the European Commission, and several International Organizations and Institutions have proposed policy strategies to reduce air pollution in cities, and its damage to health (EEA, 2022; WHO, 2021).

  

Usually, the modeling of pollutants is conducted by considering gases separately, thus failing to account for the combined effect of diverse air quality variables, or aggregating them on the basis of classical and simple techniques, sometimes based on strong assumptions (Bruno & Cocchi, 2002; K & Kumar, 2022). Air quality models can be categorized into single-pollutant and multi-pollutants models, based on the aggregation function employed (K & Kumar, 2022). Single-pollutant models identify the highest pollutant value exceeding the threshold limit, underestimating the actual air quality by neglecting the effects of other pollutants (Khanna, 2000), and the complex links among air pollutants are often neglected (Garrido et al., 2021). On the other hand, multi-pollutants based models provide an aggregated measures of air quality using various aggregation methods, such as arithmetic aggregation method (Makra, 2003), and more sophisticated methods such as fuzzy-based (Sarkheil & Rahbari, 2016), entropy function (Cheng et al., 2004), factor analysis (Bishoi et al., 2009). However, all these methodologies fail in providing the statistical assessment of their results.

Nowadays, a more reliable, extensible, and comparable index of air pollution, taking into account the complexity of the phenomenon, is mostly needed (K & Kumar, 2022; Shipley, 2016), to be employed as a policy tool for designing strategic pollution reduction programs to preserve public health.

Indeed, the aim of our paper is to propose a statistical approach to obtain an air quality index, overcoming most of the drawbacks of the currently used models.

Recent studies have emphasized the SEM's potential for modeling multidimensional phenomena. For instance, SEM is widely employed in ecology (Budtz-Jørgensen et al., 2010; Fan et al., 2016; Garrido et al., 2021), but the SEM methodology has not been completely exploited in the field of environmental research.

This paper tries to overcome this lack in the literature, proposing a unique model for different pollutants, and building a model-based multidimensional index. This new index wishes to replace the usual monitoring of single gases, and tries to overcome the usual drawbacks of environmental composite indices, for example, the strong hypotheses required and the fact that multivariate relations among contaminants are not taken into consideration. The new methodology wants to be rather general and flexible, and it can be employed with data from diverse sources. The approach used to specify the appropriate SEM for air quality modeling requires three steps: the *variable selection*, that allows to properly characterize air quality, not only on the basis of pollutants, but also by considering their determinants; the *model selection* from a set of candidate models; and the *model assessment* to evaluate the performance of the index. In this way, the methodology to obtain the new index is completely based on a statistical procedure that allows the evaluation of the best choice for each step, considering specific statistical goodness of fit criteria and significance tests.

In this study, the proposed methodology is applied to a dataset of pollutants in European Union metropolitan areas. The analyzed dataset includes Worldwide Air Quality data (https://aqicn.org/), which covers pollutants and atmospheric conditions around the world, providing unified and worldwide air quality information (Boaz et al., 2019). In addition, to take into account the features of the analyzed cities, two other sources of data at municipal level have been employed: the Organization for Economic Co-operation and Development (OECD) Metropolitan database, which provides socio-economic and environmental indicators in 36 countries (OECD, 2012), and the Eurostat metropolitan regions (NUTS3) data (Eurostat, 2019). This multiple sources, complete dataset has allowed for the accounting of the complex relationships among gases, also considering the specific social, demographic, economic and meteorological cities' features.

The most meaningful and significant exogenous manifest variables have been selected by using a Multivariate Random Forest (MRF) regression (Genuer et al., 2010; Grömping, 2009), to identify key predictors of the multivariate outcome (the six pollutants).

The latent concept of air pollution has been estimated through SEMs with different specifications, based on multiple endogenous and exogenous variables (Cole & Preacher, 2014; Fan et al., 2016; Landis et al., 2000; Tarka, 2018), identifying the best model in terms of goodness of fit. The obtained Air Pollution Index (API) has been validated by means of a sensitivity analysis that employs a simulation study to test its robustness with respect to variation of its input values to understand how variations in the underlying SEM variables affect the final index score.

Moreover, API has been compared with the currently used Air Quality Index (AQI). The AQI (Tan et al., 2021) is a well-established and largely employed tool for monitoring air pollution worldwide (Cromar et al., 2020; Xu et al., 2020). It presents the great advantage to allow the comparison in time and space, being based on the same, simple methodology employed in many different countries and for a long (period) time. The AQI constitutes an effective method to estimate air pollution levels (Suman, 2021), but it is based on debatable assumptions and thresholds, and it is built with elementary aggregation methodologies. Moreover, employing classical methodology does not benefit from the evaluation of models based on statistical fit measures and estimation of errors (Kanchan et al., 2015; Tan et al., 2021). Even if the drawbacks of AQI have been identified in previous studies (Bruno & Cocchi, 2007), the alternative proposal of

an air quality index has been based altogether on similar classical techniques, (Tan et al., 2021) even if it has been proposed quite recently. The main difference with other air quality indexes is that, while they make use of consensus-based expert opinion, the approach proposed in this paper to construct the API is empirical and model-based. However, the goal of the study is not to replace the AQI, but rather to complement it with an approach that benefits from goodness of fit measure.

Finally, to provide meaningful insights into the obtained results, a model-based cluster analysis is estimated on the municipal areas.

This paper is structured as follows: in Section 2 a description of data used for the empirical study of air quality is shown. Section 3 is devoted to the illustration of Structural Equation Models theoretical basis, with particular reference to hierarchical models with exogenous manifest variables. The empirical results are the subject of Section 4. Section 5 shows the validation of the proposed index, including a sensitivity analysis carried out by means of a simulation study, a comparison between the well-established AQI and the proposed methodology, and a model-based cluster analysis to visualize the results. Finally, in Section 6, a discussion on further developments is provided and some conclusions are drawn.

## 2 | DATA

The Worldwide Air Quality dataset is one of the main databases on air pollution. It covers around 380 main cities in the world, starting from January 2015. For each city, it includes the median for each of the air pollutant species and meteorological data, among several stations. All air pollutants are converted to the U.S. Environmental Protection Agency's (EPA) standards.

### 2.1 | Air quality data

The proposed procedure is applied considering 130 metropolitan areas across the European Union (EU). More in detail, the pollutants considered are the following. Carbon Oxide (CO) is a colorless and odorless gas formed by the incomplete reaction of air with fuel. CO pollution occurs primarily from emissions produced by fossil fuel-powered engines, including motor vehicles and non-road engines and vehicles (Stork & Anguish, 2005). Sulfur dioxide (SO2) forms when sulfur-containing fuel such as coal, oil, or diesel is burned. The largest sources of sulfur dioxide emissions are Diesel engines, electricity generation, industrial boilers, and other industrial processes such as petroleum refining and metal processing (Gad, 2005). Particle pollution (PM2.5, PM10) is a complex mixture of air-borne particles and liquid droplets composed of acids (such as nitrates and sulphates), ammonium, water, black (or "elemental") carbon, organic chemicals, metals, and soil (crustal) material (EPA, 2022; WHO, 2021). EPA groups particle pollution into two categories: "Inhalable coarse particles" such as those found near roadways and dusty industries, range in diameter from 2.5 to 10 micrometers (or microns) and "Fine particles", such as those found in smoke and haze, are 2.5 micrometers in diameter and smaller. PM2.5 is referred to as "primary" if it is directly emitted into the air as solid or liquid particles and is called "secondary" if it is formed by chemical reactions of gases in the atmosphere. Major sources of primary fine particles include cars and trucks (especially those with diesel engines), open burning, wildfires, fireplaces, woodstoves, outdoor wood boilers (also called hydronic heaters), cooking, dust from roads and construction, agricultural operations, coal, and oil-burning boilers. Main sources of secondary fine particles are power plants and some industrial processes, including oil refining and pulp and paper production. Nitrogen dioxide (NO2) is a gaseous air pollutant composed of nitrogen and oxygen and it belongs to a group of gases called nitrogen oxides. NO2 forms when fossil fuels such as coal, oil, gas, or diesel are burned at high temperatures. However, it can also form indoors when fossil fuels like wood or natural gas are burned. It is important to remark that NO2 contributes to chemical reactions that produce ozone (Mukhopadhyay & Sahu, 2018). Cars and trucks are the largest sources of emissions, followed by power plants, diesel-powered heavy construction equipment, other movable engines, and industrial boilers (EPA, 2022). Ground-level ozone (O3) is not usually directly emitted, but rather forms from chemical reactions between oxides of nitrogen and volatile organic compounds. This happens when pollutants emitted by cars, power plants, industrial boilers, refineries, chemical plants, and other sources chemically react in the presence of sunlight (Wennberg & Dabdub, 2008). For this reason, as highlighted by Mukhopadhyay and Sahu (2018), rural sites are less polluted than the urban ones in terms of NO2, whereas the converse is true for O3. This is due to the reaction with NOx emissions, which are greatest in urban areas.

## 2.2 | Air quality index

Based on the pollutants described above, the AQI is calculated as follows:

$$AQI = \max_{p} AQI_p, \qquad p = 1, \ldots, 6 \tag{1}$$

where

$$AQI_p = \frac{I_{Hi} - I_{Lo}}{BP_{Hi} - BP_{Lo}} \left(C_p - BP_{Lo}\right) + I_{Lo} \tag{2}$$

with:

- Ip = the index for pollutant p;
- Cp = the truncated concentration of pollutant p;
- BPHi, BPLo = concentration break-points greater than or equal and less than Cp, respectively;
- Ihi, Ilo = the AQI values corresponding to BPHi and BPLo, respectively.

The AQI runs from 0 to 500 (Tan et al., 2021): the higher the AQI value, the greater the level of air pollution and the greater the health concern. The AQI is divided into six categories. Each category corresponds to a different level of health concern, and it is represented by a specific color. For instance, an AQI value of 50 or below represents good air quality (green color), while an AQI value over 300 represents hazardous air quality (in brown). In general, when AQI values are above 100, air quality is unhealthy: at first for certain sensitive groups of people, then for everyone as AQI values get higher (colors from orange to red). Therefore, the AQI relies on decisions made by field experts on the break-points for AQI categories' definition.

From Figure 1, it is possible to display the distribution of air pollutants in the EU. The size of each point is proportional to the air pollutant concentration level of the city. Moreover, points are colored according to their specific AQI value. As it is possible to observe, AQIs for each pollutant vary a lot among different countries. For instance, the AQI for SO2 ranges from 0 to 30, while for CO2 and PM2.5 it goes from 0 to 200.

Broadly employed and well accepted (Cromar et al., 2020; Suman, 2021; Xu et al., 2020), AQI is based on debatable choices that may appear not fully appropriate to evaluate air quality and that may lead to some ambiguous classification (Bruno & Cocchi, 2007; Kanchan et al., 2015; Tan et al., 2021). For instance, as shown in Table 1, the "Unhealthy" category mostly includes cities with only 1 pollutant over the stated threshold. This pollutant is always PM2.5. No category includes more than 3 contaminants exceeding air quality standard in Europe.

Moreover, the first category only includes Zurich, while the last does not include any observation. This highlights a limited discriminating power of the AQI technique for the European data: the AQI considers only the pollutant with the highest concentration. However, this approach can be not exhaustive if the quite low correlations values among pollutants are considered (Supporting Information (SI), Table A). A city with only one pollutant with high concentration may belong to the same category of a city with more than one pollutant that exceeds the security threshold. For instance, consider the city of Turin (in Italy): it presents 5 pollutants over 6 with values below the air quality threshold, and only PM2.5 is in the purple category. In Milan, still in Italy and quite close to Turin, PM2.5 is in the purple category, but also NO2 and PM10 present high values. However, according to the AQI, both cities are in the purple very polluted category: this result provides some additional doubts on the cities classification according to the AQI.

## 2.3 | Exogenous manifest variables

Exogenous manifest variables (MVs), chosen based on previous studies and theoretical knowledge, have been considered to estimate the latent construct of "air pollution" to study the complex relationships among gasses and identify air pollution determinants.

The meteorological and atmospheric covariates already included in the Air Quality database have been considered. Then, the Air Quality Dataset has been merged with two databases on socio-economic data at the municipal level. The Eurostat metropolitan regions (NUTS3) database (Eurostat, 2019) provides data on the number of cars in EU cities, while

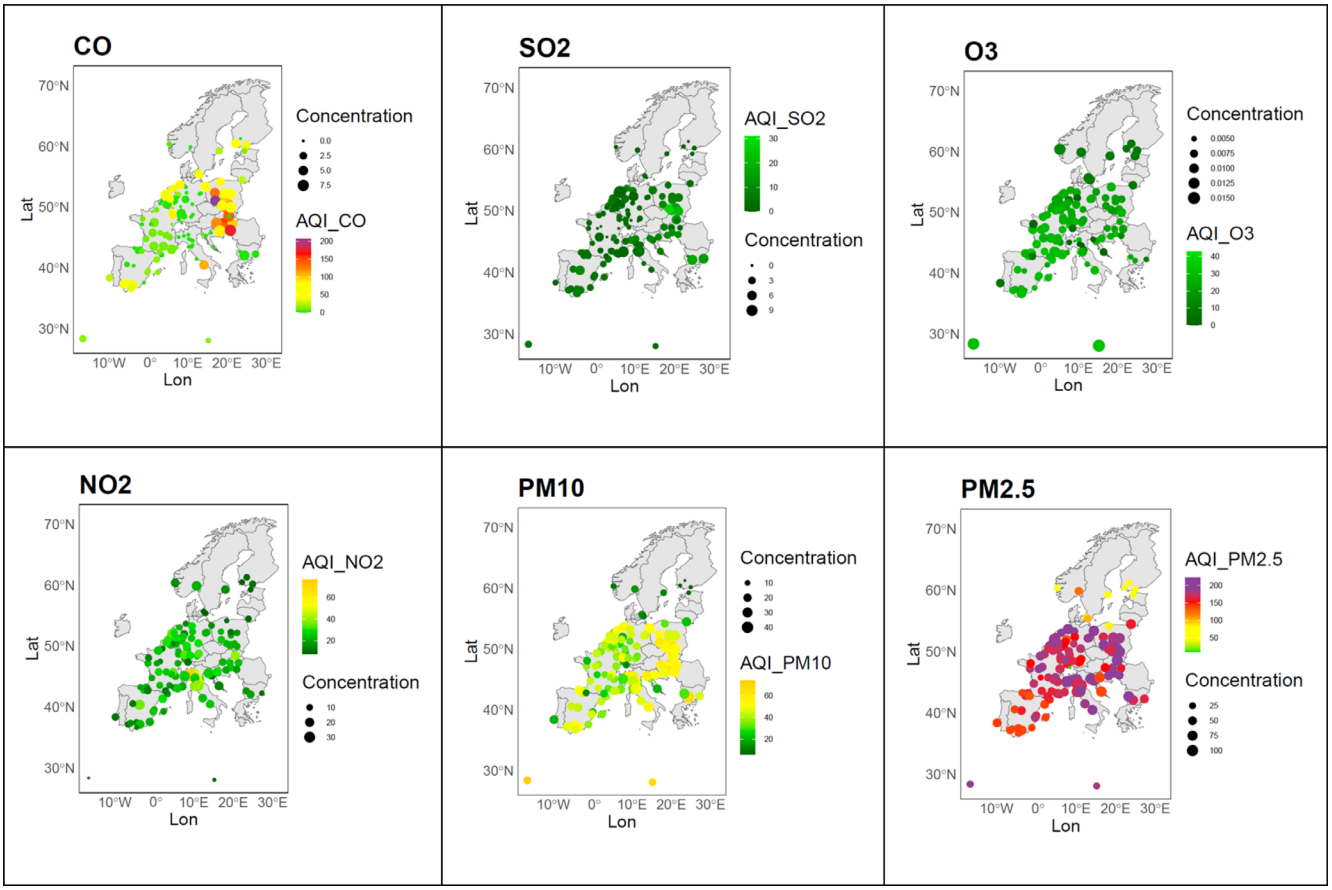


**FIGURE 1** Distribution of pollutants' concentrations for the EU cities (legend in black). Points are colored according to the corresponding pollutant-specific AQI category.

**TABLE 1** Cities exceeding air quality standards by AQI categories and number of pollutants. Percentages.

| | Number of pollutants exceeding air quality standards | | | | |
|---|---|---|---|---|---|
| **AQI categories** | **0** | **1** | **2** | **3** | **Total** |
| Good | 100 | 0 | 0 | 0 | 100 |
| Moderate | 0 | 75.0 | 25.0 | 0.0 | 100 |
| Unhealthy for sensitive | 0 | 66.7 | 33.3 | 0.0 | 100 |
| Unhealthy | 0 | 70.8 | 20.2 | 9.0 | 100 |
| Very unhealthy | 0 | 7.1 | 14.3 | 78.6 | 100 |
| Hazardous | - | - | - | - | - |
| Total | 0.8 | 63.1 | 21.5 | 14.6 | 100 |

the Metropolitan database from the OECD provides socio-economic indicators for 691 OECD functional urban areas over 250,000 inhabitants in 36 countries (OECD, 2012). The OECD and the EU have developed a harmonized definition of functional urban areas (FUAs), being composed of a city and its commuting zone. FUA's definition aims at maximizing international comparability and overcoming the limitation of using administrative aggregation approaches.

From the Air Quality database, the following variables on atmospheric conditions have been employed. Temperature (Celsius degrees): it has been studied that air temperature can affect the movement of air pollution. In fact, energy from the sun is absorbed by the Earth's surface and therefore air near the ground is warmer than the one that is in the troposphere. The warmer, lighter air at the surface rises, and the cooler, heavier air in the upper troposphere sinks. This induces the pollutants movement from the ground to higher altitudes (De Sario et al., 2013). Humidity (g/m3) affects the natural

deposition of Particulate Matter in the air. With an increase in humidity, the size of the PM also increases. It may also happen that it becomes too heavy to remain in the air and begins to fall off: this phenomenon is called dry PM deposition (Center for Science and Education, 2020). Pressure (Pa): the concentration of most air pollutants is affected by meteorological conditions, but the impact level depends on the pollutants type and varies across different areas (Liu et al., 2020). Wind-gust (m/s) and Wind-speed (m/s) may influence air pollution movements. A passing storm front can wash pollutants out of the atmosphere or transport them to a new area. In Asia, it has been noticed that powerful spring winds carry clouds of industrial pollutants from China across the Gobi Desert. These contaminated winds cross the desert, picking up particle pollution as well. This causes massive yellow dust storms across the Korean Peninsula and parts of Japan. On the other hand, the absence of wind can create stagnant air. When the air stops moving pollutants, such as vehicle and factory exhaust, concentrate over an area (Center for Science and Education, 2020).

Two geographical covariates (Latitude and Longitude) have been included in the analysis, in order to take into account the spatial configuration of the phenomenon of air pollution (Urdangarin et al., 2022). According to Liu et al. (2020), as the latitude increases, the impact of temperature on air pollutants' concentration becomes more obvious. According to Eurostat (2021), Eastern Europe has some EU's most polluted cities. In particular, the highest concentrations of dangerous fine particles are in Bulgaria (19.6 μg/m3) and Poland (19.3 μg/m3), followed by Romania (16.4 μg/m3) and Croatia (16.0 μg/m3). In contrast, the concentrations are lowest in urban areas of Estonia (4.8 μg/m3), Finland (5.1 μg/m3) and Sweden (5.8 μg/m3).

In addition, from Eurostat database, the motorization rate (Number of passenger cars per thousand inhabitants, available at country level), and the Number of registered cars per 1000 population (at city level) have been included. Traffic-related air pollutant emissions have become a global environmental problem, especially in urban areas, and recent studies point out that O3, PM2.5 and NO2 are emitted from cars' engines (Choma et al., 2021). Indeed, Eurostat's road transport emission data are used in several studies on air pollution. For instance, Antanasijević et al. (2018) employs a multiple-input–multiple-output general regression neural network model, based on basic socioeconomic and transport related indicators for the simultaneous prediction of sulfur oxides, nitrogen oxides, ammonia, non-methane volatile organic compounds and PM emissions at the national level.

Vegetation and green spaces have shown reductive effects on air-borne pollutants concentrations, especially of PM (Diener & Mudu, 2021; Villani et al., 2021). Therefore, also the "Share of land (%): Green urban areas and sports and leisure facilities" Eurostat indicator has been included. Scientific assessments of agricultural air quality are an important emerging area of environmental science that offers significant challenges to policy and regulatory authorities. Agricultural emissions play an important role in several atmospherically mediated processes of environmental and public health concerns. These atmospheric processes affect local and regional environmental quality, including PM exposure (Aneja et al., 2009). To take into account possible emissions due to agricultural activities, the Eurostat "Share of land (%): Agricultural areas" indicator at city level has been considered in the analysis.

Recent literature has focused on the links between pollution and socio-economic characteristics (Martori et al., 2022; Ren & Matsumoto, 2020). Therefore, socio-economic and demographic variables have been included in the dataset for each European city. These variables come from the OECD database and refer to population features in urban areas, such as GDP per capita, population density, and employment rate at city level. For instance, an OECD study (Dechezleprêtre et al., 2020) provides evidence that air pollution causes economy-wide reductions in market economic activity based on data for Europe. An instrumental variables approach based on thermal inversions is used to identify the causal impact of air pollution on economic activity. The estimates show that a 1 μg/m3 increase in PM2.5 concentration (or a 10% increase at the sample mean) causes a 0.8% reduction in real GDP that same year. Ninety-five per cent of this impact is due to reductions in output per worker, which can occur through greater absenteeism at work or reduced labour productivity. Therefore, results suggest that public policies to reduce air pollution may contribute positively to economic growth. Indeed, the large economic benefits from pollution reduction uncovered in the study compare with relatively small abatement costs. Literature on the relationships between GDP and pollution is very wide and presents mixed results. It is not the aim of this paper to state the sign and intensity of such link. However, in this analysis, the GDP Per Capita (GDPPC in Million USD, constant prices, constant PPP, base year 2015) at city level, has been included, representing the economic development level as one of the major driving forces of air pollution in large cities (Chen & Kan, 2008). High population density (PD, inhabitants per $km^2$) is expected to induce an increasing in pollution emissions, especially for developing countries (Chen & Kan, 2008). According to Borck and Schrauth (2019), based on a panel analysis for Germany, a one-standard deviation increase in population density increases air pollution by 3%–12%. Other socio-economic and demographic variables considered in literature are the Elderly Dependency Ratio (EDR), the Employment and Unemployment Rate. For instance, the unemployment rate has an expected negative impact on all pollutants, especially on CO and NO2 (Davis, 2012).

## 3 | STATISTICAL FRAMEWORK

Structural equation modeling enables researchers to simultaneously model and estimate complex relationships among multiple dependent and independent variables. The concepts under consideration are typically unobservable and modeled indirectly by multiple indicators. In estimating the relationships, SEM accounts for measurement error in observed variables. As a result, a model for the theoretical concepts of interest is obtained (Cole & Preacher, 2014). The SEM methodology has seen a huge attention recently for its wide application possibilities and its powerful estimation capabilities (Budtz-Jørgensen et al., 2010; Hair et al., 2019; Hair et al., 2020).

The model-based approaches allow to split the relations among latent, manifest endogenous and exogenous variables in two parts (Landis et al., 2000). First, there is a structural model that links together the latent endogenous and exogenous variables corresponding to latent constructs, usually represented by ovals in the path diagram (Hair et al., 2020). Second, there are the measurement models of the latent variables that display the relationships between the constructs and the manifest variables, represented by rectangles in the path diagram (Figure 2).

The classical SEM formulation is reported (3), $i$ refers to the generic population unit (Table 2).

$$\begin{cases} \eta_i = \mathbf{B}\eta_i + \boldsymbol{\Gamma}'\xi_i + \zeta_i & (\textit{Structural model}) \\ y_i = \mathbf{C}\eta_i + \varepsilon_i & (\textit{Measurement model for endo variables}) \\ x_i = \mathbf{A}\xi_i + \delta_i & (\textit{Measurement model for exo variables}) \end{cases} \tag{3}$$

Now, let us suppose that a sample of size $n$ from the population of $N$ units is drawn. Thus, SEM can be formulated in a compact matrix form. The structural model and the measurement models of endogenous and exogenous variables are presented in (4), (5) and (6), respectively.

$$\boldsymbol{N} = \boldsymbol{N}\boldsymbol{B}' + \boldsymbol{F}\boldsymbol{\Gamma} + \boldsymbol{Z} \tag{4}$$

$$\boldsymbol{Y} = \boldsymbol{N}\boldsymbol{C}' + \boldsymbol{E} \tag{5}$$

$$\boldsymbol{X} = \boldsymbol{F}\boldsymbol{A}' + \boldsymbol{D} \tag{6}$$

Path models are diagrams used to visually display the hypotheses and variable relationships that are examined when SEM is applied (Figure 2). The left part shows the measurement model for the exogenous latent variables in Equation (6) (i.e., those constructs that only explain other constructs in the model). In the right part, the measurement model for the endogenous LVs in Equation (5) (i.e., those constructs that are being explained in the model) is depicted. In this

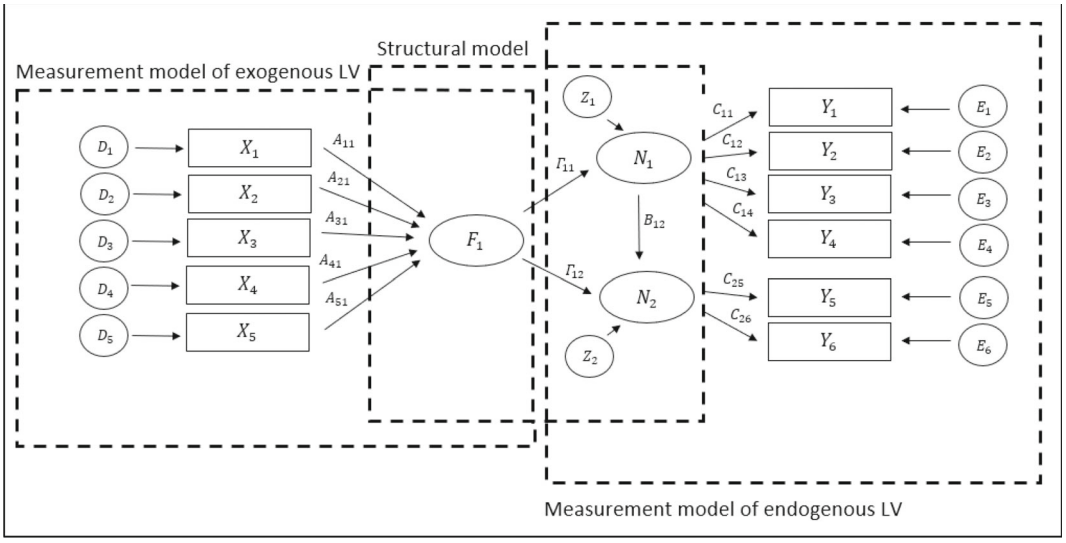


FIGURE 2 SEM path model.

**TABLE 2** SEM vector and matrix notations.

| Vector/matrix | Coefficients |
|---|---|
| $\eta_i$ $(L \times 1)$ | Factor scores of the endogenous latent variables (LVs) for unit $i$ |
| B $(L \times L)$ | Regression coefficients of the endogenous LV, $L$ is the number of endogenous LVs |
| $\Gamma$ $(H \times L)$ | Regression coefficients of exogenous LVs, $H$ is the number of exogenous LVs |
| $\xi_i$ $(H \times 1)$ | Ractor scores of the exogenous LVs for unit $i$ |
| $\zeta_i$ $(L \times 1)$ | Errors of the structural model associated to unit $i$ |
| $y_i$ $(M \times 1)$ | Endogenous Manifest Variables (MVs) for unit $i$, $M$ is the number of endogenous MVs |
| C $(M \times L)$ | Coefficients (covariances or correlations) between MVs and endogenous LVs (orthogonal matrix) |
| $\varepsilon_i$ $(M \times 1)$ | Errors of the measurement model for endogenous LVs associated to unit $i$ |
| $x_i$ $(J \times 1)$ | Exogenous MVs for unit $i$, $J$ is the number of exogenous MVs |
| A $(J \times H)$ | Coefficients (covariances or correlations) between exogenous MVs and LVs, $J$ is the number of exogenous MVs (orthogonal matrix) |
| $\delta_i$ $(J \times 1)$ | Errors of the measurement model for exogenous LVs |
| N $(n \times L)$ | Factor scores of endogenous LVs |
| F $(n \times H)$ | Factor scores of exogenous LVs |
| Z $(n \times L)$ | Errors of the structural model |
| Y $(n \times M)$ | Endogenous MVs |
| E $(n \times M)$ | Errors of the measurement model of endogenous variables |
| X $(n \times J)$ | Exogenous MVs |
| D $(n \times J)$ | Errors of the measurement model for exogenous variables |

measurement part, the relations between MVs and LVs are studied. Finally, the central part of the path diagram contains the structural part, the one that studies the relationships of the latent constructs among themselves (4). All relations are estimated simultaneously (Hair & Sarstedt, 2021; Tarka, 2018).

To estimate SEM coefficients, the formulation of the model covariance matrices should be investigated. The theoretical covariance matrices in (7)–(11) can be obtained extending the sample to all the population of $N$ units:

$$\begin{aligned}
\boldsymbol{\Sigma}_{\mathbf{N}} &= \left(\frac{\mathbf{1}}{\boldsymbol{N}}\right)\mathbf{N}'\mathbf{N} = \left(\frac{\mathbf{1}}{\boldsymbol{N}}\right)\left(\left(\mathbf{F}\boldsymbol{\Gamma}' + \mathbf{Z}\right)\left(\mathbf{I} - \mathbf{B}'\right)^{-\mathbf{1}}\right)'\left(\left(\mathbf{F}\boldsymbol{\Gamma}' + \mathbf{Z}\right)\left(\mathbf{I} - \mathbf{B}'\right)^{-\mathbf{1}}\right) = \\
&= (\mathbf{I} - \mathbf{B})^{-\mathbf{1}}\left(\left(\frac{\mathbf{1}}{\boldsymbol{N}}\right)\left(\mathbf{F}\boldsymbol{\Gamma}' + \mathbf{Z}\right)'\left(\mathbf{F}\boldsymbol{\Gamma}' + \mathbf{Z}\right)\right)\left(\mathbf{I} - \mathbf{B}'\right)^{-\mathbf{1}} = \\
&= (\mathbf{I} - \mathbf{B})^{-\mathbf{1}}\left(\left(\frac{\mathbf{1}}{\boldsymbol{N}}\right)\boldsymbol{\Gamma}\mathbf{F}'\mathbf{F}\boldsymbol{\Gamma} + \left(\frac{\mathbf{1}}{\boldsymbol{N}}\right)\mathbf{Z}'\mathbf{Z} + \boldsymbol{\Gamma}\left(\frac{\mathbf{1}}{\boldsymbol{N}}\right)\mathbf{F}'\mathbf{Z} + \left(\frac{\mathbf{1}}{\boldsymbol{N}}\right)\mathbf{Z}'\mathbf{F}\boldsymbol{\Gamma}'\right)\left(\mathbf{I} - \mathbf{B}'\right)^{-\mathbf{1}} = \\
&= (\mathbf{I} - \mathbf{B})^{-\mathbf{1}}\left(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{F}}\boldsymbol{\Gamma}' + \boldsymbol{\Sigma}_{\mathbf{Z}}\right)\left(\mathbf{I} - \mathbf{B}'\right)^{-\mathbf{1}}
\end{aligned} \tag{7}$$

Supposing that $\mathrm{COV}(\mathbf{F}, \mathbf{Z}) = \left(\frac{1}{N}\right)\mathbf{F}'\mathbf{Z} = 0$.

The variance covariance matrix of $\mathbf{Y}$ can be rewritten as:

$$\boldsymbol{\Sigma}_{\mathbf{YY}} = \left(\frac{1}{N}\right)\mathbf{Y}'\mathbf{Y} = \left(\frac{1}{N}\right)\mathbf{C}\mathbf{N}'\mathbf{N}\mathbf{C} + \left(\frac{1}{N}\right)\mathbf{C}\mathbf{N}' + \mathbf{E} + \left(\frac{1}{N}\right)\mathbf{E}'\mathbf{N}\mathbf{C}' = \left(\mathbf{C}\boldsymbol{\Sigma}_{\mathbf{N}}\mathbf{C}' + \boldsymbol{\Sigma}_{\mathbf{E}}\right) \tag{8}$$

Supposing that $\mathrm{COV}(\mathbf{N}, \mathbf{E}) = \left(\frac{1}{N}\right)\mathbf{N}'\mathbf{E} = 0$, we have:

$$\boldsymbol{\Sigma}_{\mathbf{YY}} = \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{N}}\mathbf{C}' + \boldsymbol{\Sigma}_{\mathbf{E}} = \mathbf{C}\left[(\mathbf{I} - \mathbf{B})^{-1}\left(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{F}}\boldsymbol{\Gamma}' + \boldsymbol{\Sigma}_{\mathbf{Z}}\right)\left(\mathbf{I} - \mathbf{B}'\right)^{-1}\left(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{F}}\boldsymbol{\Gamma}' + \boldsymbol{\Sigma}_{\mathbf{Z}}\right)\left(\mathbf{I} - \mathbf{B}'\right)^{-1}\right]\mathbf{C}' + \boldsymbol{\Sigma}_{\mathbf{E}} \tag{9}$$

The variance covariance matrix of $\mathbf{X}$ can be written as:

$$\boldsymbol{\Sigma}_{\mathbf{XX}} = \left(\frac{1}{N}\right)\mathbf{X}'\mathbf{X} = \mathbf{A}\left(\frac{1}{N}\right)\mathbf{F}'\mathbf{F}\mathbf{A}' + \left(\frac{1}{N}\right)\mathbf{D}'\mathbf{D} + \mathbf{A}\left(\frac{1}{N}\right)\mathbf{F}'\mathbf{D} + \left(\frac{1}{N}\right)\mathbf{D}'\mathbf{F}\mathbf{A}' = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{F}}\mathbf{A}' + \boldsymbol{\Sigma}_{\mathbf{D}} \tag{10}$$

Supposing that $\mathrm{COV}(\mathbf{F}, \mathbf{D}) = \left(\frac{1}{N}\right)\mathbf{F}'\mathbf{D} = 0$ a non-orthogonal FA model is considered. Similarly:

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathbf{XY}} &= \left(\frac{1}{N}\right)\mathbf{X}'\mathbf{Y} = \left(\frac{1}{N}\right)\left(\mathbf{F}\mathbf{A}' + \mathbf{D}\right)'\left(\mathbf{N}\mathbf{C}' + \mathbf{E}\right) = \left(\frac{1}{N}\right)\mathbf{A}\mathbf{F}'\mathbf{N}\mathbf{C}' + \left(\frac{1}{N}\right)\mathbf{D}'\mathbf{E} + \left(\frac{1}{N}\right)\mathbf{A}\mathbf{F}'\mathbf{E} + \left(\frac{1}{N}\right)\mathbf{E}'\mathbf{F}\mathbf{A}' \\ &= \left(\frac{1}{N}\right)\mathbf{A}\mathbf{F}'\mathbf{N}\mathbf{C}' + 0 + 0 + 0 = \left(\frac{1}{N}\right)\mathbf{A}\mathbf{F}'\left(\mathbf{F}\boldsymbol{\Gamma}' + \mathbf{Z}\right)\left(\mathbf{I} - \mathbf{B}'\right)^{-1}\mathbf{C}' = \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{F}}\boldsymbol{\Gamma}'\left(\mathbf{I} - \mathbf{B}'\right)^{-1}\mathbf{C}'\end{aligned} \tag{11}$$

If the structural equation is included in the endogenous measurement model equation, we obtain:

$$\mathbf{Y} = \left(\mathbf{F}\boldsymbol{\Gamma}' + \mathbf{Z}\right)\left(\mathbf{I} - \mathbf{B}'\right)^{-1}\mathbf{C}' + \mathbf{E} = \mathbf{F}\boldsymbol{\Gamma}'^{\left(\mathbf{I}-\mathbf{B}'\right)^{-1}}\mathbf{C}' + \mathbf{Z}\left(\mathbf{I} - \mathbf{B}'\right)^{-1}\mathbf{C}' + \mathbf{E} \tag{12}$$

Therefore, the SEM can be rewritten starting from the joint distribution of the manifest variables $\mathbf{X}, \mathbf{Y}$ in a block data matrix $[\mathbf{X}, \mathbf{Y}]$:

$$[\mathbf{X}, \mathbf{Y}] = \left[\mathbf{F}\mathbf{A}',\ \ \mathbf{F}\boldsymbol{\Gamma}'^{\left(\mathbf{I}-\mathbf{B}'\right)^{-1}\mathbf{C}'}\right] + \left[\mathbf{D},\ \ \mathbf{Z}\left(\mathbf{I} - \mathbf{B}'\right)^{-1}\mathbf{C}' + \mathbf{E}\right] \tag{13}$$

Thus, to summarize, the SEM requires the following hypotheses:

1. $\mathrm{E}(\mathbf{X}) = \mathrm{E}(\mathbf{Y}) = \mathrm{E}(\mathbf{N}) = \mathrm{E}(\mathbf{F}) = \mathrm{E}(\mathbf{E}) = \mathrm{E}(\mathbf{Z}) = \mathrm{E}(\mathbf{D}) = 0$, (all centred variables);
2. $\mathrm{COV}(\mathbf{F}, \mathbf{Z}) = \mathrm{COV}(\mathbf{N}, \mathbf{E}) = \mathrm{COV}(\mathbf{D}, \mathbf{F}) = \mathrm{COV}(\mathbf{N}, \mathbf{D}) = \mathrm{COV}(\mathbf{F}, \mathbf{E}) = 0$, (uncorrelated LVs and errors);
3. $\mathrm{COV}(\mathbf{Z}, \mathbf{E}) = \mathrm{COV}(\mathbf{Z}, \mathbf{D}) = \mathrm{COV}(\mathbf{E}, \mathbf{D}) = 0$, (null correlation between errors);
4. $(\mathbf{I} - \mathbf{B})$ is of maximum rank;
5. All variables are supposed normally distributed.

Now we can rewrite the variance covariance matrix of the entire SEM model as

$$\begin{aligned}\boldsymbol{\Sigma} &= \left(\frac{1}{N}\right)[\mathbf{X},\ \ \mathbf{Y}]'[\mathbf{X}, \mathbf{Y}] = \\ &= \left(\frac{1}{N}\right)\left(\left[\mathbf{F}\mathbf{A}',\ \ \mathbf{F}\boldsymbol{\Gamma}'^{\left(\mathbf{I}-\mathbf{B}'\right)^{-1}\mathbf{C}'}\right]\right)'\left(\left[\mathbf{F}\mathbf{A}',\ \ \mathbf{F}\boldsymbol{\Gamma}'^{\left(\mathbf{I}-\mathbf{B}'\right)^{-1}\mathbf{C}'}\right]\right) \\ &\quad + \left(\frac{1}{N}\right)\left(\left[\mathbf{D},\ \ \mathbf{Z}\left(\mathbf{I} - \mathbf{B}'\right)^{-1}\mathbf{C}' + \mathbf{E}\right]\right)'\left[\mathbf{D},\ \ \mathbf{Z}\left(\mathbf{I} - \mathbf{B}'\right)^{-1}\mathbf{C}' + \mathbf{E}\right] = \\ &= \begin{bmatrix} \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{F}}\mathbf{A}' & \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{F}}\boldsymbol{\Gamma}'\left(\mathbf{I} - \mathbf{B}'\right)^{-1}\mathbf{C}' \\ \mathbf{C}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{F}}\mathbf{A}' & \mathbf{C}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{F}}\boldsymbol{\Gamma}'^{\left(\mathbf{I}-\mathbf{B}'\right)^{-1}\mathbf{C}'} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{D}} & 0 \\ 0 & \mathbf{C}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Sigma}_{\mathbf{Z}}\left(\mathbf{I} - \mathbf{B}'\right)^{-1}\mathbf{C}' + \boldsymbol{\Sigma}_{\mathbf{E}} \end{bmatrix} = \\ &= \begin{bmatrix} \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{F}}\mathbf{A}' + \boldsymbol{\Sigma}_{\mathbf{D}} & \mathbf{A}\boldsymbol{\Sigma}_{\mathbf{F}}\boldsymbol{\Gamma}'\left(\mathbf{I} - \mathbf{B}'\right)^{-1}\mathbf{C}' \\ \mathbf{C}(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{F}}\mathbf{A}' & \mathbf{C}(\mathbf{I} - \mathbf{B})^{-1}\left(\boldsymbol{\Gamma}\boldsymbol{\Sigma}_{\mathbf{F}}\boldsymbol{\Gamma}' + \boldsymbol{\Sigma}_{\mathbf{Z}}\right)\left(\mathbf{I} - \mathbf{B}'\right)^{-1}\mathbf{C}' + \boldsymbol{\Sigma}_{\mathbf{E}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{XX}} & \boldsymbol{\Sigma}_{\mathbf{XY}} \\ \boldsymbol{\Sigma}_{\mathbf{YX}} & \boldsymbol{\Sigma}_{\mathbf{YY}} \end{bmatrix}\end{aligned} \tag{14}$$

Recalling that the product between LVs and errors and between errors are for hypothesis null. $\boldsymbol{\Sigma}_{\mathbf{XX}}$, $\boldsymbol{\Sigma}_{\mathbf{XY}}$,$\boldsymbol{\Sigma}_{\mathbf{YY}}$ are the variance and covariance matrices of the measurement and structural models.

Let us define $\mathbf{S}$ the sample variance and covariance matrix of the MVs $[\mathbf{X},\ \ \mathbf{Y}]$ :

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{\mathbf{XX}} & \mathbf{S}_{\mathbf{XY}} \\ \mathbf{S}_{\mathbf{YX}} & \mathbf{S}_{\mathbf{YY}} \end{bmatrix} \text{ and recall the estimators } \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{XX}} & \boldsymbol{\Sigma}_{\mathbf{XY}} \\ \boldsymbol{\Sigma}_{\mathbf{YX}} & \boldsymbol{\Sigma}_{\mathbf{YY}} \end{bmatrix}. \tag{15}$$

Different estimation methods for the SEM model can be adopted: Least Square (LS), Maximum Likelihood (ML), and Generalized LS. In this study, a ML estimation has been employed and the discrepancy function to minimize is:

$$F_{ML}[\boldsymbol{\Sigma}] = \log|\boldsymbol{\Sigma}| + tr\left[\mathbf{S}\boldsymbol{\Sigma}^{-1}\right] - \log|\mathbf{S}| - (J + M) \tag{16}$$

where $(J + M)$is the number of MVs. The Likelihood is distributed according to $\chi^2$with degree of freedom: $df = \frac{1}{2}(J + M)(M + J + 1) - t$where $t$ is the number of estimated parameters.

## 4 | EMPIRICAL RESULTS

SEM models are implemented using "sem" function of the R "lavaan" package (Rosseel, 2012). This R function automatically standardized the six pollutants, assigning negative weights to the variables that the factor reconstructs in the opposite direction from the others.

The process of modifying the model to improve its fit to the empirical data seems to be necessary for all SEM models (Tarka, 2018). In this study, a forward specification was employed, starting with a baseline model, and then estimating two more comprehensive models (Bentler & Chou, 1987), exploiting the results provided by an Exploratory Factor Analysis (EFA) and considering the additional information provided by the available exogenous MVs. Following this forward approach, the first model is an initial Confirmatory Factor Analysis model with six indicators of a single latent variable (measurement model). The second model (Section 4.2) introduces a *hierarchy*, considering two latent factors, and subsequently a structural model among the latent constructs. This approach is based on the results on an EFA estimated on the six pollutants (SI, Table B) and the internal consistency analysis on the one factor SEM (Section 4.1). In the third model, a measurement model of exogenous MVs is introduced to exploit the information available from the explanatory covariates (Section 2.3).

The estimated models' fit has been compared based on several indices (Section 4.4). In this way, it is possible to verify if a very simple baseline model is sufficient to model simultaneously the six pollutants in the best way or if a more complex hierarchical model is needed to provide more accurate modeling of the phenomenon. The comparison with the third model assesses the significance of the selected exogenous MVs to model city air pollution.

### 4.1 | Single factor model estimation

The first SEM specification aimed at modeling city air pollution, based on the six different pollutants concentrations, and obtaining an Air Pollution Index (API_1), foresees a single latent construct, directly related to the six pollutants. This baseline specification allows to estimate the API_1 as a weighted combination of the pollutants' concentration, considering directly the endogenous MVs coefficients as weights (Figure 3a).

All six pollutants coefficients' estimates are strongly significant ($p$ values of the significance tests always equal to zero). The model estimates also the variance for each pollutant. All path coefficients are in modulus greater than 0.3, meaning that the "air pollution" single factor reconstructs well the 6 pollutants, in particular the 2 PM ones. The model correctly detects that O3 is reconstructed in the opposite direction with respect to the other 5 pollutants.

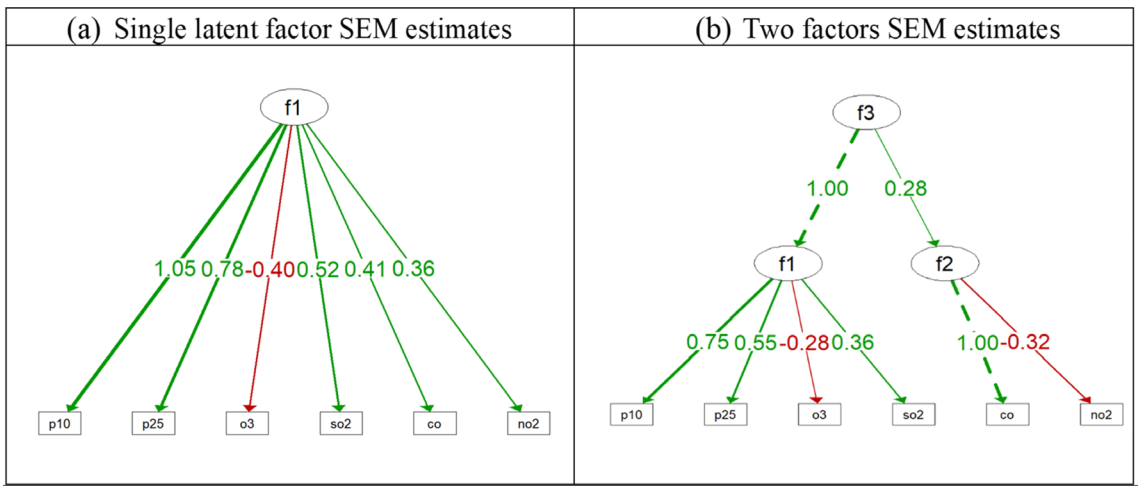


FIGURE 3 SEM estimates. All latent factors are standardized.

This model presents some advantages: it is easy to understand and it allows the direct calculation of the API_1. The relation between pollutants and API_1 is straightforward: policy makers know that, reducing for example, PM10, the overall pollution will be reduced by the amount indicated by the estimate coefficient (weight). However, it presents some limitations: it does not consider the cities' features and therefore it does not exploit the information power of some explanatory variables that could be considered to assess the quality of the air.

## 4.2 | Hierarchical path-model estimation

Hierarchical models measure latent concepts defining sets of MVs. Therefore, the general latent concept can be represented by a tree structure where each internal node represents a specific order of abstraction for the latent concept measured (Cavicchia & Vichi, 2022). In this study, the higher order general latent factor, representing a broader concept of "air pollution", is estimated through a Higher Order SEM, taking into account two latent concepts, grouping different pollutants (as in the measurement model on the right of Figure 2). The air pollutants modeling the two latent factors are chosen based on the EFA presented in the SI. The system of equations simultaneously estimated is:

$$\begin{cases} f1 = PM10 + PM2.5 + O3 + SO2 \\ f2 = CO + NO2 \\ f3 = f1 + f2 \end{cases} \tag{17}$$

The 2 first-order factors and the second-order one are estimated simultaneously and produce an estimate of the latent concept "air pollution" called API_2_s.

Some constraints for the hierarchical path model are set to fulfill the restrictions on the degrees of freedom required for the SEM estimation. In fact, to allow estimates' convergence, some constrains must be introduced. For instance, when a single factor loads on two observable variables (e.g., CO and NO2), the parameters to estimate are 2 (one for each manifest indicator), but the information provided is just 1, namely the covariance between the 2 indicators. Consequently, the degree of freedom is equal to −1. The solution usually applied is to constrain the loadings to be equal to 1 and/or constrain some variances to be unitary: in this way, the degree of freedom is 0 and the model can be identified.

Therefore, the following model constraints had to be introduced: (1) The path coefficient of CO in the second factor is set equal to 1. The choice of fixing to 1 the coefficient of CO and not NO2 is due to the fact that CO has a positive relationship with the latent construct, while NO2 has a negative coefficient. (2) The path coefficient of f1 is set equal to 1. In this case, the choice of constraining the coefficient of f1 and not f2 is because f1 loads on more pollutants than f2 and has a larger impact on the higher order latent construct f3. (3) To allow the convergence of the estimates of the impact of NO2 and CO on f3, their variances are set equal to 1.

Figure 3b reports model's estimates together with their standard errors. According to the usual R notation of packages "lavaan" and "semPlot", the equal sign of Equation (17) is graphically represented as arrows pointing outwards. All loadings and variances estimates are strongly significant. The 2 first order factors (f1 and f2) loadings present the same sign, with the first showing a triple coefficient, compared to the second. Each pollutant's importance in determining the API is similar to the one estimated in the single factor SEM (Section 4.1). PM10 presents the highest coefficient, followed by PM2.5, confirming the role of PM in causing city pollution.

## 4.3 | Hierarchical path-model with explanatory variables estimation

Given the constraints of the traditional SEM approaches, the predictors (or exogenous MVs) are usually selected on the basis of theoretical or empirical considerations. To optimally select the explanatory variables to be employed in the estimation of the latent concept of air pollution, based on a SEM, the MRF regression seemed the most appropriate approach, since it provides information about the importance of each MV, taking into account simultaneously all the six pollutants (Grömping, 2009; Jacobucci et al., 2019; Speiser, 2022). Random forests, introduced by Breiman (2001), are a very efficient methodology for both classification and regression. The MRF regression takes as input all the possible explanatory variables to predict the 6 pollutants variables simultaneously. The most meaningful and appropriate exogenous MVs have been identified following a backward approach, correcting any possible errors of inclusion. To perform variable selection,

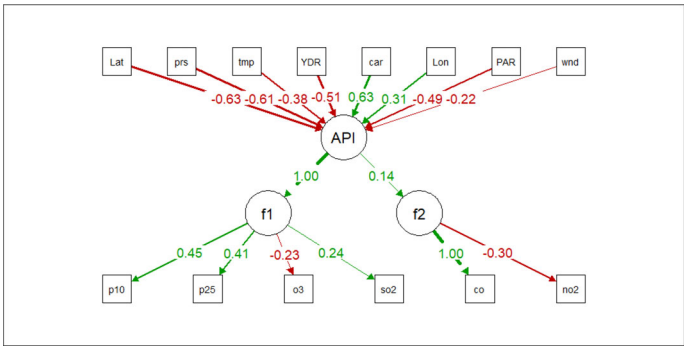


**FIGURE 4** Two factors with exogenous MVs SEM estimates. All latent factors are standardized.

a score of importance of each covariate must be assigned. In the regression random forests framework, the most widely used importance score for a given variable is the increasing in Mean Square Error (MSE) of a tree in the forest, when the observed values of this variable are randomly permuted in the Out Of Bag (OOB) samples (Genuer et al., 2010). Another importance score is based on the Mahalanobis distance. It has the advantage of being additive in the components of the outcomes and it does not consider correlations between the continuous coordinates of the outcome vector (Ishwaran & Kogalur, 2022).

In the analysis, the 12 most important variables have been chosen, corresponding to those that, included in the tree, produced the highest reductions in the MSE. Most of them remain the same according also to the Mahalanobis criterion. In the latter, the employment rate variable appears in the first 12 places (SI, Table C).

Among these variables, the number of circulating cars appears to be, by far, the most important according to the MSE criterion, and one of the most important for the Mahalanobis one. The number of new cars (registered in the year of reference), at city level, also appears important for the level of air pollution, being the fifth selected variable with the MSE method, and the eighth on the basis of the Mahalanobis distance. Other variables such as latitude, longitude, temperature, air pressure and humidity are identified as important drivers of air pollution. Economic and social features of European cities contribute to the air pollution estimates: the share of young population, the characteristics of the labour market in terms of unemployment and participation rates and the level of GDP per capita all appear among the most important variables, according to the MRF methodology. On the other hand, the estimates indicate that the share of agricultural areas and green spaces are not among the variables most influencing urban air pollution.

Moreover, starting from a more general SEM, including all the 12 selected covariates, a backward procedure has been applied to the explanatory variables, to further select the most meaningful ones, used to estimate the API_2 (SI, Table D). More in detail, the coefficients related to total number of passenger cars, Elderly Dependency Ratio, GDP per capita, and Humidity are not significant. Therefore, the final SEM is estimated considering 8 variables.

The final model can written as follows:

$$\begin{cases} f1 = PM2.5 + PM10 + O3 + SO2 \\ f2 = CO + NO2 \\ API = f1 + f2 \\ API \sim Lat + Lon + car + tmp + prs + YDR + PAR + wnd \end{cases} \tag{18}$$

The estimated relationships between the 6 pollutants and the 2 latent concepts related to "air pollution" can be graphically seen in the path diagram below (Figure 4), while the estimated standard errors for all the variables included in the model are reported in SI, Table D.

The equal sign of Equation (18) is graphically represented as arrows pointing outwards and the tilde sign as arrows pointing inwards, referring to the exogenous MVs (as in Figure 2). Results show that loadings related to the six pollutants are greater or close to the threshold of 0.3 and strongly significant (all $p$-values are very low). The first factor (f1) includes the same four pollutants as the previous model (Section 4.2). Also in this case, PM10 and PM2.5 present the highest coefficients, followed by SO2.

The number of cars in the city (car) strongly influences the whole air pollution level: an increase of this variable by 1, implies an API_2 value 0.63 higher. Northern cities are less likely to be polluted (Latitude (Lat)

**TABLE 3** Effect of each pollutant on API_2.

| Pollutant | Effect on API_2 |
|---|---|
| PM10 | 0.45 |
| PM2.5 | 0.41 |
| O3 | −0.23 |
| SO2 | 0.24 |
| CO | 0.14 |
| NO2 | −0.04 |

coefficient equal to −0.63), while Eastern European countries present a higher probability of being polluted (Longitude (Lon) coefficient of 0.31). The Youth Dependency ratio (YDR: population under 15 years of age, compared to the number of people of working age, i.e., 15–64) presents a negative coefficient, indicating that "younger" cities are also less polluted. Negative effects on the air pollution level are also given by the Participation to the labour market (PAR) coefficient. Both temperature (tmp) and pressure (prs) coefficients are negative: these results are in line with what is known in literature on the subject (see Section 2.3).

Results obtained analyzing the proposed API_2 measure are in line with literature on the subject that takes into account single pollutants. The advantages of this methodology are that it exploits the additional information provided by exogenous MVs and directly explains how the determinants of the air pollution influence the total API_2. In other words, the exogenous variables can be considered as drivers of the pollution concentration and therefore they can be useful to policy makers. The fact that all the coefficients of the exogenous MVs are significant means that inclusions of such variables in the model is indeed useful for a better estimate.

Table 3 shows the effects of each pollutant on the API_2; they are obtained as in Bollen (2011) considering the coefficients of Figure 4 (e.g., the PM10 coefficient is obtained multiplying 0.45, which is the effect of PM10 on f1 by 1, which is the effect of f1 on API_2).

The highest effects for API_2 are those of Particulate Matter emissions, which are the same 2 pollutants that determine most of the AQI values for the 130 cities. However, while -for instance- the AQI assigns effect equal to 1 to a single pollutant and 0 to all the others, the proposed index attributes a specific effect to each pollutant, considering also the negative effects that some pollutants may have on the air pollution index (for example, O3).

## 4.4 | Evaluation of the estimated models

In the final step of the model assessment procedure, models' performances have been compared. The most appropriate SEM to describe air pollution is chosen considering the evaluation of the path coefficients' significance and the analysis of the model's explanatory and predictive power (Shmueli et al., 2019). In research situations, when a theory is not well established or confirmed by the empirical observation, the comparison of alternative models is required. To compare different model configurations and select the best model, an information criterion likelihood-based finding a trade-off between the fit of the model and its complexity (specifically the number of parameters) should be used. With this aim, the Bayesian Information Criterion (BIC) is employed. The model, which yields the smallest BIC value, is considered the best model in the set. In addition, the Akaike Information Criterion (AIC) can be chosen to offer further evidence for a model compared to alternative models in the set (Hair & Sarstedt, 2021).

The three theoretically justifiable competing models (API_1, API_2_s and API_2), presented in the previous section are here compared based on the most reliable measures of goodness of fit and information criteria (Table 4).

AIC and BIC facilitate the comparison of models in terms of model fit (Burnham & Anderson, 2004; Hair & Sarstedt, 2021), providing a relative likelihood of being the data generation model, given the data and a set of competing models (Danks et al., 2020).

The Root Mean Square Error of Approximation (RMSEA) and the Standard Root Mean Squared Residual (SRMR) also provide measure of errors of the models (Pavlov et al., 2021).

From the overall analysis, the best measure appears to be the API_2 (Table 4).

**TABLE 4** Goodness of fit measures for the three models.

| | API_1 | API_2_s | API_2 |
|---|---|---|---|
| AIC | 1954.139 | 2010.268 | 1969.935 |
| BIC | 1988.549 | 2038.943 | 1979.551 |
| RMSEA | 0.255 | 0.306 | 0.176 |
| SRMR | 0.146 | 0.316 | 0.178 |

## 5 | VALIDATION OF THE SEM-BASED AIR POLLUTION INDEX

Indexes have a crucial role, functioning not only as descriptive tools but also as aids that can impact significant policies. Even if the proposed index is based on a statistical procedure that benefits intrinsically from the assessment of its goodness of fit, we rigorously verify the index construction, employing simulations to perform a sensitivity analysis of the latent construct (Grimaccia et al., 2023; Janováa et al., 2019; Ozenne et al., 2022).

Once the best SEM specification to describe air pollution has been chosen (Section 4.4), next step is the model results' validation. Results will show that the proposed index represents accurately the concept of air pollution, as intended in the literature.

A sensitivity analysis is carried out to monitor the projected behavior of the new index in 3 possible scenarios. The simulation relies on the observation of the changes in the model behavior upon variation of its input values.

Furthermore, a graphical visualization of the solutions, as well as a cluster analysis, are being used to provide a better understanding of the index.

Finally, an external validation has been carried out comparing the model's results with those of the well validated AQI, considering data for the years 2019 and 2022.

### 5.1 | Sensitivity analysis of API_2 and AQI: Simulation study

The sensitivity analysis is conducted through a simulation study that examines, for API_2 and AQI, three different scenarios: Scenario A, where pollution improves over time with decreasing emissions; Scenario B, where pollution remains stable with constant emissions; and Scenario C, where pollution worsens over time with increasing emissions. Each scenario involves running 1000 simulations every 3 months from January 2023 to December 2030.

In Scenario A, starting from January 2023, the emissions for each subsequent period $(t+1)$ are obtained by subtracting a percentage of the emission at period t from the emission at period t. To determine the distribution of this percentage, the variation of the 6 pollutants over the 36-month period from 2019 to 2022 has been studied. For example, PM2.5 increased on average by 30% in three years. Assuming a linear increase over time, the increase every 3 months is calculated using the geometric mean. To compute the geometric mean (Equation 19), an initial value of 100% and a final value of $100\% + 30\% = 130\%$ are considered.

$$\text{Geometric Mean} = \left(\frac{\textit{Final value}}{\textit{Initial value}}\right)^{\left(\frac{1}{\text{Number of 3-Month Periods}}\right)} - 1 \tag{19}$$

In this case, there are 12 three-month periods in 3 years, resulting in a geometric mean of approximately 0.0242. Therefore, the percentage used to perturbate the value of PM2.5 will follow a normal distribution with a mean of 0.0242 and a standard deviation equal to twice the mean. The same approach is applied to the other five pollutants, each with their respective means (CO: 0.0376, SO2: 0.0241, O3: 0.0196, PM10: 0.0201, NO2: 0.0242).

Scenario C is obtained in a similar manner, but instead of subtracting the percentage, it is added. Finally, in Scenario B, a normally distributed percentage with mean 0 and standard deviation of 0.01 is added. To allow comparisons between the AQI and API_2 indexes, both are normalized in 0–1. The AQI takes values in 0–500 and therefore its range is 500. The range is divided into 6 categories, corresponding each to the 10%, 10%, 10%, 10%, 20% and 40% of the overall range: Good (green) Moderate (yellow), Unhealthy for Sensitive Groups (orange), Unhealthy (red), Very Unhealthy (purple) and Hazardous (brown), as indicated by the experts' recommendations. When we normalize the AQI to the 0–1 range, the breakpoints for each category are correspondingly rescaled to 0.1, 0.2, 0.3, 0.4, and 0.6.

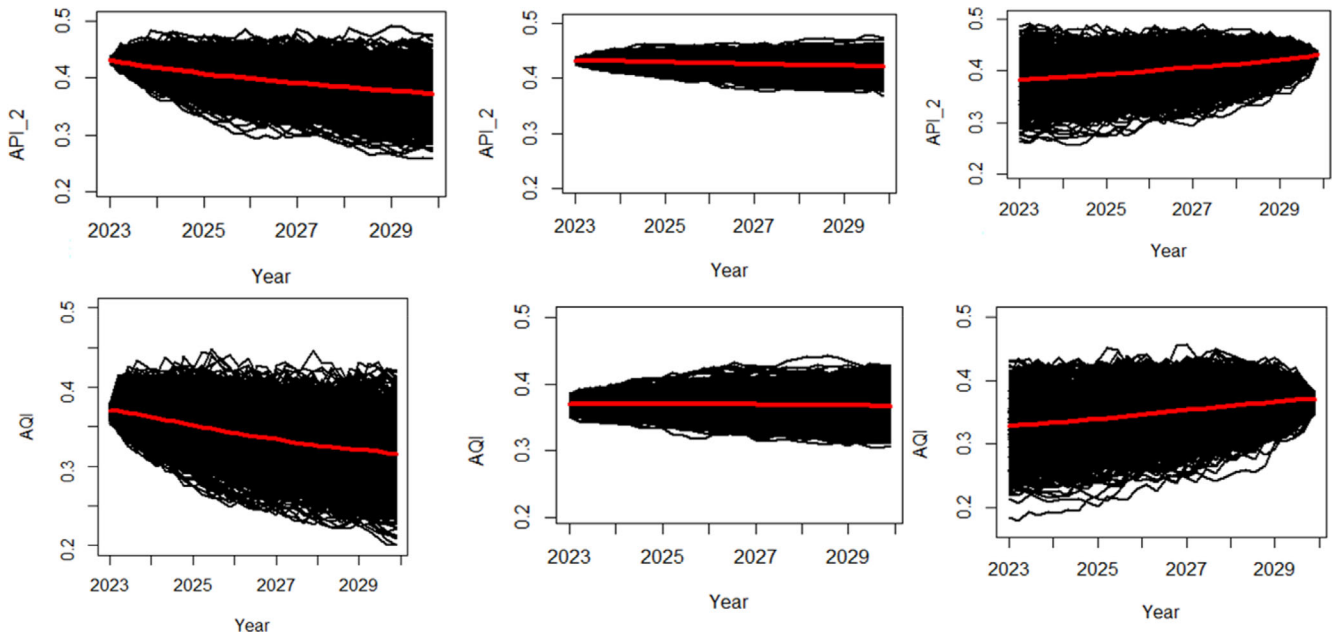


**FIGURE 5** Top-left: simulation A, evolution of API_2 with slight pollutants' reduction. Top-center: simulation B, evolution of the API_2 with constant values of the pollutants. Top-right: simulation C, evolution of the API_2 with slight pollutants' increase. Bottom-left: simulation A for AQI. Bottom-center: simulation B for AQI. Bottom-right: simulation C for AQI.

Similarly, we have applied the same normalization approach to the new index. To determine the range of API_2, we consider the index value when all pollutants are set to 0, as well as when they assume the breakpoints' values corresponding to the highest AQI category. This method, based on the E.P.A. concentrations' quality standards (refer to SI, Table E), SI ensures a consistent and comparable scale for both indices. By employing this normalization technique, we can visually compare AQI and API_2 (Figures 6 and 7). Figure 5 depicts the results for the three scenarios for both indexes, showing the mean API_2 and AQI over the 1000 simulations highlighted in red, for each of the 32 periods considered.

The simulation results demonstrate API_2's stability under perturbations in pollutant levels and its accurate representation of input value variations. Its mean values are all slightly higher with respect to the AQI means. Notably, API_2 has a lower dispersion around the mean with respect to AQI: this uncertainty is of approximately 5% for Scenarios A and C and even smaller for Scenario B, while this percentage increases to 10% considering AQI. This reduced variability further underscores the API_2 precision and reliability.

## 5.2 | Graphical comparison of AQI with the new index API_2

The analysis of the two indexes' distributions for the European cities adds information on the quality of their classifications (Figure 6). The histograms' columns are colored according to the categories to which cities belong. API_2 shows a more symmetrical distribution, with the 50% of the cities below 0.28. Considering AQI, most of the cities are concentrated in the red Unhealthy category. This is due mainly to the PM2.5 that almost everywhere determines the AQI (70.8% of Unhealthy and 7.1% of Very Unhealthy). It is possible to conclude that the API_2 better discriminates among different air pollution values, compared to the AQI: firstly, the presence of cities in the sixth Hazardous category (a category in which AQI does not have any city in); and secondly, the larger number of cities falling within the first "Good" category.

This is because API_2 considers the effect of each pollutant and does not solely rely on the one of the pollutant with the maximum emission level, as the AQI does.

Figure 7 shows the geographical distribution of cities according to the six categories for both indexes. Considering the example for Turin and Milan (see Section 2.2), the API_2 index correctly identifies that Milan is more polluted with respect to Turin and places Milan in the purple Very Unhealthy category and Turin in the orange Unhealthy for sensitive groups category.

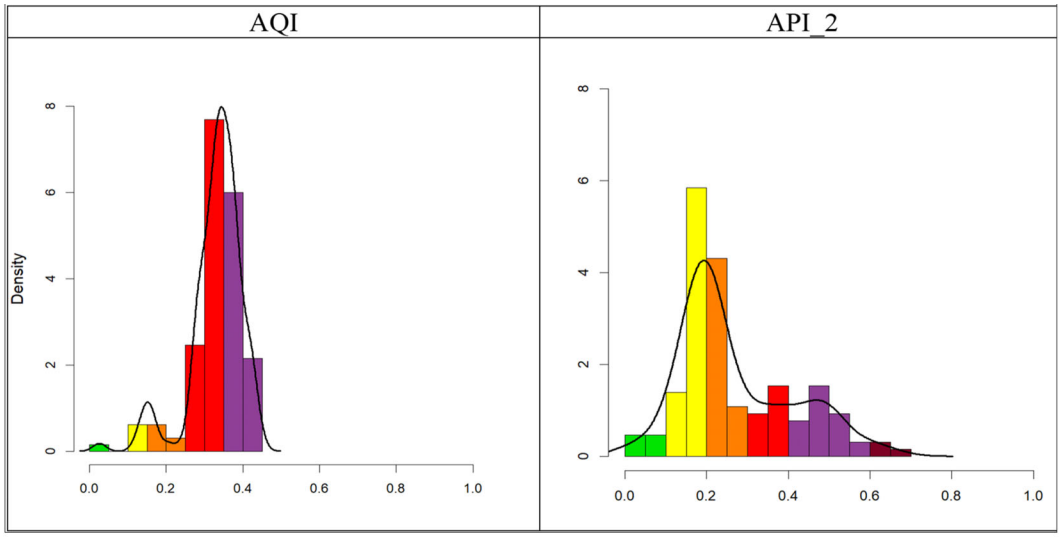


**FIGURE 6** Distribution of AQI and API_2 values for the European cities in 2019. Both indexes are normalized in 0–1.

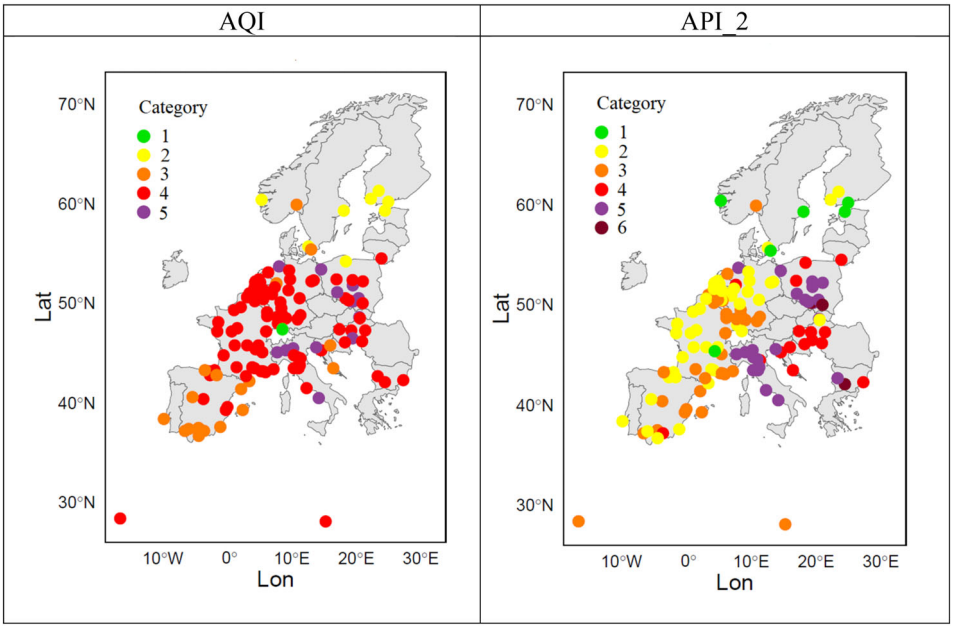


**FIGURE 7** AQI and API_2 distributions (normalized in 0–1) for the 130 EU cities (year 2019). Points are colored according to the corresponding category.

According to both indexes, the less polluted cities are in Finland and Estonia. On the other hand, the most polluted areas are in the North of Italy, Hungary, Poland and Bulgaria.

Almost all French and Spanish cities are "Moderate" for the API_2 and "Unhealthy" for the AQI.

Finally, the less polluted cities (yellow points) have mostly lower longitude with respect to the "Unhealthy" Eastern cities. These results demonstrate that API_2 validity is supported also by its alignment with existing literature and other air quality indexes.

## 5.3 | Cluster analysis

An interesting possibility offered by the proposed approach is to cluster cities on the basis of the new index of air pollution (Bottazzi et al., 2023). The 130 European cities are grouped into clusters, homogeneous with respect to the air pollution level, each represented by a centroid that corresponds to an API_2 value. The clustering technique of k-means is applied.

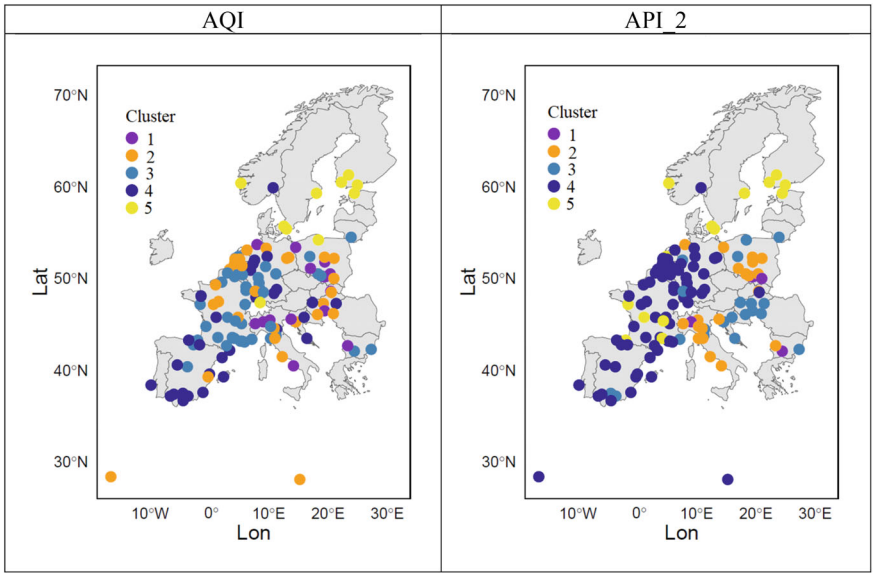


**FIGURE 8** Clusters of cities with respect to homogeneous air pollution level (year 2019). The optimal number of clusters has been chosen according to the maximum Average Silhouette Width.

Groups are ranked from 1 to k (k = total number of clusters) considering the centroids' values from the highest to the lowest: rank 1 corresponds to the largest centroid and therefore to the group of more air polluted cities. In this study, according to the silhouette average method, the optimal number of clusters is 5 for both indexes (SI, Figure 1).

The grouping of cities is shown in Figure 8. These maps show the air pollution distribution in Europe, highlighting groups of cities with a similar situation in terms of air pollution levels.

The clustering obtained employing the API_2 presents less points in the "average" group, and more often assigns the same ranking to cities in the same country.

The meaningful clustering of units based on API_2 highlights its capability to provide valuable insights into air quality categorization.

## 5.4 | Comparison of API_2 and AQI for 2022 data

To assess the external consistency of the innovative index, data for 2022 have been considered. The air pollution's distributions in 2022 are analyzed, considering both the currently used AQI and the proposed API_2 (first 2 plots, Figure 9). Furthermore, clusters of cities in 2022 are compared (last 2 plots, Figure 9).

Data from 2022 is available for 79 European cities only; because of this, the following analysis is conducted just on the 70 cities that provided data for both years.

To compute the API_2 for 2022, the coefficients estimated for 2019 are employed as weights for 2022 values of the six air pollutants.

From 2019 to 2022, for both indexes, the overall air pollution level increased. Being the API_2 really multidimensional, it accounts for the increasing of all the pollutants and not only of the most widespread as PM2.5 and PM10. Consequently, API_2 demonstrates a more evident reflection of this increasing effect, as depicted in Figure 9 (first 2 plots on the left). The successful detection for API_2 of the increase in air pollution that occurred between 2019 and 2022 reinforces its reliability as a robust index. Cluster analysis, proposed in Section 5.3, has been estimated for the AQI and API_2 referring to 2022. Results show that the classification of cities according to the new API_2 in 2022 is still meaningful: 5 groups of cities with respect to air pollution level have been identified (always considering the maximum average silhouette width), with Cluster 1 containing the most polluted cities, while Cluster 5 the less polluted ones.

It is possible to see that Cluster 1 contains the same cities for both indexes, Clusters 2 and 3 have a higher cardinality for AQI while the cardinality of Clusters 4 and 5 is higher considering API_2.

The higher discriminative power of API_2 is further confirmed in the second two plots of Figure 9. Specifically, the new index allocates fewer units to Cluster 3, which contains cities with an average air pollution level. This outcome highlights the ability of API_2 to categorize air quality levels more effectively than AQI.

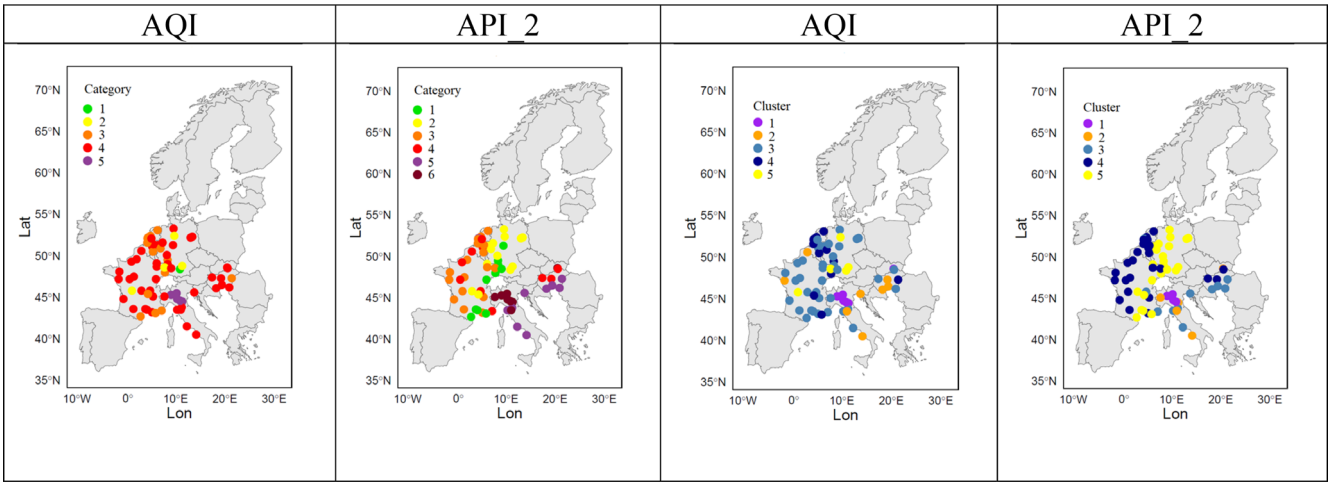


**FIGURE 9** First 2 maps on the left: distributions for the 70 EU cities (year 2022). Points are colored according to the corresponding category. Second 2 maps on the right: clusters of cities with respect to homogeneous air pollution level (year 2022).

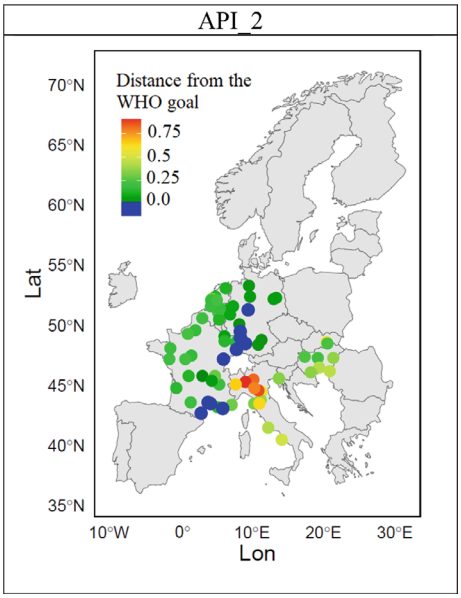


**FIGURE 10** Distance of each city from the 2030 air pollution WHO goal in terms of API_2 value (range 0–1). The blue dots represent cities which already met the target.

To conclude, the obtained results for the API_2 are in line with previous knowledge about air pollution, from literature and from other indices. The innovation in the proposed methodology is the SEM based approach and the statistical procedure that relies each choice at every step on statistical measures of significance and goodness of fit. This procedure overcomes subjective and debatable choices in the index definition.

## 5.5 | The new index (API_2) as a decision support tool

The novel index holds a great potential as a decision support tool, empowering the implementation of sustainable policies at both national and EU levels. More specifically, API_2 can be employed to draw policy scenarios. For instance, based on the WHO guidelines (WHO, 2021),[1] the maximum acceptable levels of pollutants correspond to an API_2 value of 0.117. The mean API_2 value for the European cities is 0.424 in 2022.

By examining the index for the 130 European cities in 2019, those represented with blue dots have already met the target identified by WHO for 2030 (Figure 10).

Assuming a 15% annual decrease from 2022 and 2030, according to the projected API_2 value, approximately the 68% of the European cities will reach the WHO target.

## 6 | DISCUSSION AND CONCLUSIONS

An innovative procedure for developing an Air Pollution Index, based on an optimal specification of a Structural Equation Model (SEM), is proposed in this paper. The added value of our approach is in its reliance on objective choices rather than subjective thresholds and legislative assumptions made by field specialists. At each step of the procedure, we meticulously assess the goodness of fit and significance of the model, ensuring a rigorous and robust methodology and accounting for possible error estimation. Moreover, a significant advantage of the proposed methodology is the identification of exogenous manifest variables that are meaningful drivers of air pollution and that can be employed as policy making tools to reduce air pollution.

Indeed, many air pollution models lacks in the interpretation of the synergistic and antagonistic relationships among pollutants (Khanna, 2000). Moreover, multi-pollutants models encounter rigidity issues, as the employed aggregation functions may not be adaptable to the inclusion of new pollutants (Makra, 2003). For instance, fuzzy based information models can be not unequivocally interpletable, since they require a priori knowledge to be provided by field experts (Sarkheil & Rahbari, 2016). In contrast, SEMs offer a more reliable approach, allowing for the straightforward incorporation of all relevant parameters: as a result, the final estimates automatically adjust to accommodate the additional parameters, providing a deep understanding of the determinants influencing the API_2. Moreover, the selection of parameters is guided by statistical goodness-of-fit measures.

Another drawback of traditional multi-pollutants models is the choice of the aggregation function that could overestimate or underestimate pollutants' effects. SEMs address this concern by offering a model-based aggregation of multiple determinants, mitigating subjectivity and concerns related to over or underestimation.

To our knowledge, the index discussed in this paper is the first SEM-based index of air pollution that simultaneously considers the six main contaminant gases and significant pollution drivers.

Indeed, applying the proposed methodology to the European cities, model-based weights for each pollutant on the latent concept of air pollution have been estimated. The results show that PM10 and PM2.5 have the strongest impact (weight) on the Air Pollution Index, while Ozone exhibits a negative impact (consistently with literature). The most meaningful explanatory variable, that can be used as leverage to reduce air pollution in European cities, is the number of cars. The meteorological conditions of a city are also key aspects to take into account: in cities with higher temperature and pressure, the air pollution level is lower on average, the other covariates fixed.

Although the robust statistical procedure proposed benefits by itself of statistically measured optimal choices, the new index has been further validated. A sensitivity analysis assesses its robustness with respect to variation of its input values and a face validation is used to confirm that the index's output is in line with existing knowledge on air pollution in Europe. Finally, the external validity of API_2 is checked assessing the coherence of its results with the currently used AQI. API_2 has great advantages with respect to AQI, including its totally objective construction approach and its higher discriminating power among different levels of air pollution. This superiority arises from the fact that API_2 employs a statistical model that considers simultaneously the effect of each pollutant, rather than only the one with the highest emission.

In conclusion, the paper proposes a statistical methodology to model air pollution, employing a hierarchical SEM with exogenous manifest variables, selected through a multivariate random forest regression. However, the proposed procedure remains rather flexible, as it can be applied to derive an index measuring different phenomena, by considering the optimal specification of a SEM. Each step of the index construction process relies on statistical choices made considering the goodness of fit as well as the significance of the obtained results.

### DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in Air Quality Historical Data Platform at https://aqicn.org/. These data were derived from the following resources available in the public domain: - Air Quality Historical Data Platform, https://aqicn.org/data-platform/register/

## ENDNOTE

[1] Annual average concentrations of $PM_{2.5}$ should not exceed 5 μg/m$^3$, while 24-h average exposures should not exceed 15 μg/m$^3$. $PM_{10}$ (particulate matter with a diameter of 10 microns or less) concentrations of 15 μg/m$^3$ annual mean, 45 μg/m$^3$ 24-h mean. Ozone ($O_3$) concentrations of 100 μg/m$^3$ 8-h mean. Nitrogen dioxide ($NO_2$) concentrations of 10 μg/m$^3$ annual average and 25 μg/m$^3$ 24-h mean. Sulfur dioxide ($SO_2$) concentrations of 40 μg/m$^3$ 24-h mean. Carbon monoxide (CO) concentrations of 7 μg/m$^3$ 24-h mean.

## ORCID

*Mariaelena Bottazzi Schenone* https://orcid.org/0000-0002-8905-2389
*Elena Grimaccia* https://orcid.org/0000-0001-6816-0666
*Maurizio Vichi* https://orcid.org/0000-0002-3876-444X

## REFERENCES

Aneja, V. P., Schlesinger, W. H., & Erismanal, J. W. (2009). Effects of agriculture upon the air quality and climate: Research, policy, and regulations. *Environmental Science & Technology*, *43*, 4234–4240.

Antanasijević, D., Pocajt, V., Peric-Grujic, A., & Ristic, M. (2018). Multiple-input–multiple-output general regression neural networks model for the simultaneous estimation of traffic-related air pollutant emissions. *Atmospheric Pollution Research*, *9*, 388–397.

Bentler, P. M., & Chou, C. P. (1987). Practical issues in structural equation modeling. *Sociological Methods & Research*, *16*(1), 78–117.

Bishoi, B., Prakash, A., & Jain, V. K. (2009). A comparative study of air quality index based on factor analysis and US-EPA methods for an urban environment. *Aerosol and Air Quality Research*, *9*(1), 1–17.

Boaz, R. M., Lawson, A. B., & Pearce, J. L. (2019). Multivariate air pollution prediction modelling with partial missingness. *Environmetrics*, *30*(7). https://doi.org/10.1002/env.2592

Bollen, K. A. (2011). Evaluating effect, composite, and causal indicators in structural equation models. *MIS Quarterly*, *35*(2), 359–372.

Borck, R., & Schrauth, P. (2019). Population density and urban air quality. *Regional Science and Urban Economics*, *86*(c). https://doi.org/10.1016/j.regsciurbeco.2020.103596

Bottazzi Schenone, M., Grimaccia, E., & Vichi, M. (2023). *Optimal number of clusters to rank a model-based index*. In Vichi, Zaccaria, & Mingione (Eds.), *High-quality and timely statistics—new methods and applications*. Springer nature in press.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Bruno, F., & Cocchi, D. (2002). A unified strategy for building simple air quality indices. *Environmetrics*, *13*, 243–261.

Bruno, F., & Cocchi, D. (2007). Recovering information from synthetic air quality indices. *Environmetrics*, *18*, 345–359.

Budtz-Jørgensen, E., Debes, F., Weihe, P., & Grandjean, P. (2010). Structural equation models for meta-analysis in environmental risk assessment. *Environmetrics*, *21*(5), 510–527.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304.

Cavicchia, C., & Vichi, M. (2022). Second-order disjoint factor analysis. *Psychometrika*, *87*(1), 289–309.

Center for Science Education. (2020). How weather affects air quality. Center for Science Education. https://scied.ucar.edu/learning-zone/air-quality/how-weather-affects-air-quality

Chen, B., & Kan, H. (2008). Air pollution and population health: A global challenge. *Environmental Health and Preventive Medicine*, *13*(2), 94–101.

Cheng, W. L., Kuo, Y. C., Lin, P. L., Chang, K. H., Chen, Y. S., Lin, T. M., & Huang, R. (2004). Revised air quality index derived from an entropy function. *Atmospheric Environment*, *38*(3), 383–391.

Choma, E. F., Evansb, J. S., Gomez-Ibanezc, J. A., Did, Q., Schwartzb, J. D., Hammitte, J. K., & Spenglerb, J. D. (2021). Health benefits of decreases in on-road transportation emissions in the United States from 2008 to 2017. *PNAS*, *118*(51). https://doi.org/10.1073/pnas.2107402118

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300–315.

Cromar, K. R., Ghazipura, M., Gladson, L. A., & Perlmutt, L. (2020). Evaluating the U.S. air quality index as a risk communication tool: Comparing associations of index values with respiratory morbidity among adults in California. *PLoS One*, *15*(11), e0242031.

Danks, N. P., Pratyush, N., Sharma, B., & Sarstedt, M. (2020). Model selection uncertainty and multimodel inference in partial least squares structural equation modeling (PLS-SEM). *Journal of Business Research*, *113*, 13–24.

Davis, M. E. (2012). Recessions and health: The impact of economic trends on air pollution in California. *American Journal of Public Health*, *102*(10), 1951–1956.

De Sario, M., Katsouyanni, K., & Michelozzi, P. (2013). Climate change, extreme weather events, air pollution and respiratory health in Europe. *European Respiratory Journal*, *42*, 826–843.

Dechezleprêtre, A., Rivers, N., & Stadler, B. (2020). The economic cost of air pollution: Evidence from Europe. In *OECD Economics Department Working Papers*, 1584. OECD Publishing. https://doi.org/10.1787/56119490-en

Diener, A., & Mudu, P. (2021). How can vegetation protect us from air pollution? A critical review on green spaces' mitigation abilities for air-borne particles from a public health perspective—with implications for urban planning. *Science of the Total Environment*, *796*, 148605.

Dominici, F., Samet, J. M., & Zegeral, S. L. (2000). Combining evidence of air pollution and daily mortality from the 20 largest US cities: A hierarchical modelling strategy. *Journal of the Royal Statistical Society, Series A*, *163*(3), 263–302.

EEA. (2022). Air quality in Europe 2022 Report no. 05/2022. https://doi.org/10.2800/488115

EPA. (2022). U.S. Transportation Sector Greenhouse Gas Emissions 1990–2020. Fast Facts. Office of Transportation and Air Quality. EPA-420-F-22-018.

Eurostat. (2019). Methodological manual on territorial typologies. Luxembourg. https://doi.org/10.2785/930137

Eurostat. (2021). How polluted is the air in urban areas? EDN-20210603-1.

Fan, Y., Chen, J., Shirkey, G., Ranjeet, J., Wu, R., Park, H., & Shao, C. (2016). Applications of structural equation modeling (SEM) in ecological studies: An updated review. *Ecological Processes*, *5*(19). https://doi.org/10.1186/s13717-016-0063-3

Gad, S. C. (2005). *Sulfur Dioxide*. In P. Wexler (Ed.), *Encyclopedia of toxicology* (2nd ed., pp. 115–116). Elsevier.

Garrido, M., Hansen, S. K., Yaari, R., & Hawlena, H. (2021). A model selection approach to structural equation modelling: A critical evaluation and a road map for egologists. *Methods in Ecology and Evolution*, *13*, 42–53.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236.

Grimaccia, E., Naccarato, A., & Solari, F. (2023). Computational procedures for quality assessment of latent concepts. *Stat*, *12*(1), e569. https://doi.org/10.1002/sta4.569

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random Forest. *The American Statistician*, *63*(4), 308–319.

Hair, J. F., Babin, B. J., Anderson, R. E., & Black, W. C. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.

Hair, J. F., & Sarstedt, M. (2021). Explanation plus prediction – The logical focus of project management research. *Project Management Journal*, *52*(4), 319–322.

Hair, J. F., Wolfinbarger, C. M., Money, A. H., Samouel, P., & Page, M. J. (2020). *Essentials of business research methods* (4th ed.). Routledge.

Hoskovec, L., Martenies, S., Burket, T. L., Magzamen, S., & Wilson, A. (2022). Association between air pollution and COVID-19 disease severity via Bayesian multinomial logistic regression with partially missing outcomes. *Environmetrics*, *33*(1). https://doi.org/10.1002/env.2751

Ishwaran, H., & Kogalur, U. B. (2022). RandomForestSRC: Multivariate splitting rule.

Jacobucci, R., Brandmaier, A. M., & Kievit, R. A. (2019). A practical guide to variable selection in structural equation modeling by using regularized multiple-indicators, multiple-causes models. *Advances in Methods and Practices in Psychological Science*, *2*(1), 55–76.

Janováa, J., Hampel, D., & Nerudováb, D. (2019). Design and validation of a tax sustainability index. *European Journal of Operational Research*, *278*, 916–926.

Kanchan, K., Gorai, A. K., & Goyal, P. (2015). A review on air quality indexing system. *Asian Journal of Atmospheric Environment*, *9*(3), 101–113.

Khanna, N. (2000). Measuring environmental quality: An index of pollution. *Ecological Economics*, *35*(2), 191–202.

K, P., & Kumar, P. (2022). A critical evaluation of air quality index models (1960–2021). *Environmental Monitoring and Assessment*, *194*, 324.

Landis, R. S., Beal, D. J., & Tesluk, P. E. (2000). Comparison of approaches to forming composite measures in structural equation models. *Organizational Research Methods*, *3*(2), 186–207.

Liu, Y., Zhou, Y., & Lu, J. (2020). Exploring the relationship between air pollution and meteorological conditions in China under environmental governance. *Scientific Reports*, *10*, 14518.

Makra, L. (2003). Evaluation of the air quality of Szeged with some assessment methods. *Acta Climatologica*, *36*, 85–92.

Martori, J. C., Lagonigro, R., & Pascual, R. I. (2022). Sustainable cities and society social status and air quality in Barcelona: A socio-ecological approach. *Sustainable Cities and Society*, *87*, 104210.

Mukhopadhyay, S., & Sahu, S. K. (2018). A Bayesian spatiotemporal model to estimate long-term exposure to outdoor air pollution at coarser administrative geographies in England and Wales. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *181*(2), 465–486.

OECD. (2012). *Redefining "urban". A new way to measure metropolitan areas*. OECD Publishing. https://doi.org/10.1787/9789264174108-en

Ozenne, B., Budtz-Jørgensen, E., & Ebert, S. E. (2022). Controlling the familywise error rate when performing multiple comparisons in a linear latent variable model. *Computational Statistics*, *38*, 1–23.

Pavlov, G., Maydeu, O. A., & Shi, D. (2021). Using the standardized root mean squared residual (SRMR) to assess exact fit in structural equation models. *Educational and Psychological Measurement*, *81*(1), 110–130.

Ren, L., & Matsumoto, K. (2020). Effects of socioeconomic and natural factors on air pollution in China: A spatial panel data analysis. *Science of the Total Environment*, *740*, 140155.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1-36.

Sarkheil, H., & Rahbari, S. (2016). Development of case historical logical air quality indices via fuzzy mathematics (Mamdani and Takagi-Sugeno systems), a case study for Shahre Rey town. *Environmental Earth Sciences*, *75*(19), 1–13.

Shipley, B. (2016). *Cause and correlation in biology: A user's guide to path analysis, structural equations and causal inference with R* (2nd ed.). Cambridge University Press.

Shmueli, G., Sarstedt, M., Hair, J. F., Cheah, J., Ting, H., Vaithilingam, S., & Ringle, C. M. (2019). Predictive model assessment in PLS-SEM: Guidelines for using PLSpredict. *European Journal of Marketing*, *53*(11), 2322–2347.

Speiser, J. L. (2022). A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data. *Journal of Biomedical Informatics*, *117*, 103763.

Stork, C., & Anguish, D. (2005). *Carbon Monoxide*. In P. Wexler (Ed.), *Encyclopedia of toxicology* (2nd ed., pp. 423–425). Elsevier.

Suman. (2021). Air quality indices: A review of methods to interpret air quality status. *Materials Today Proceedings*, *34*, 863–868.

Tan, X., Han, L., Zhang, X., Zhou, W., Li, W., & Qian, Y. (2021). A review of current air quality indexes and improvements under the multi-contaminant air pollution exposure. *Journal of Environmental Management*, *279*, 111681.

Tarka, P. (2018). An overview of structural equation modeling: Its beginnings, historical development, usefulness and controversies in the social sciences. *Quality & Quantity*, *52*, 313–354.

Urdangarin, A., Goicoa, T., & Ugarte, M. D. (2022). Evaluating recent methods to overcome spatial confounding. *Revista Matemática Complutense*, *36*, 333–360.

Villani, M. G., Russo, F., Adani, M., Piersanti, A., Vitali, L., Tinarelli, G., Ciancarella, L., Zanini, G., Donateo, A., & Rinaldi, M. (2021). Evaluating the impact of a wall-type green infrastructure on PM10 and NOx concentrations in an urban street environment. *Atmosphere*, *12*, 839.

Wennberg, P. O., & Dabdub, D. (2008). Rethinking ozone production. *Science*, *319*, 1624–1625.

World Health Organization. (2021). *WHO global air quality guidelines. Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. World Health Organization.

Xu, K., Cui, K., Young, L. H., Wang, Y. F., Hsieh, Y. K., Shun, W. S., & Jiajia, Z. J. (2020). Air quality index, indicatory air pollutants and impact of COVID-19 event on the air quality near Central China. *Aerosol and Air Quality Research*, *20*, 1204–1221.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Bottazzi Schenone, M., Grimaccia, E., & Vichi, M. (2024). Structural equation models for simultaneous modeling of air pollutants. *Environmetrics*, e2837. https://doi.org/10.1002/env.2837