# Statistical methods for evaluating the environmental sustainability

Andrea Gilardi[1] and Francesca Ieva[1]

MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133

**Abstract.** This paper represents a preliminary attempt at exploring a series of statistical methods for evaluating the environmental sustainability of urban and rural territories. We briefly present and discuss modern statistical techniques to study the relationships between a set of air pollutants. In particular, we focus on multivariate methods to explore air quality data after encoding the atmospheric measurements as covariance matrices that summarise the relationships among pollutants at different monitoring sites. In fact, a key property of covariance matrices is that they lie on a Riemannian manifold, and we exploit this fact to facilitate the exploratory analyses. Future directions of our work include extending the methods discussed here in a geostatistical setting, employing techniques such as Kriging for Riemannian data.

**Keywords:** air quality, geostatistical models, kriging, riemannian manifolds

## 1 Introduction

In recent years, air pollution is emerging as a pressing global and national concern, especially in some areas of the country near the Po Valley [4]. Recognizing this challenge, the European Commission recently stated that national and supranational policies should integrate climate change as a priority area to facilitate the transition towards a low-carbon economy. Therefore, during the last few years, the European Union (EU) promoted several projects targeting the environmental sector and, more precisely, the sustainability transition. These initiatives encompass diverse plans such as waste collection, development of renewable energy sources, or, as discussed here, air quality improvement.

Exploratory analyses, dimensionality reduction methods, and spatio-temporal air quality models are key to develop ex-ante forecasting or ex-post evaluation of the impact of the EU investments mentioned before, playing also a pivotal role to assess whether a project improved (or, at least, influenced) the environmental sustainability of a territory.

The starting point of the aforementioned techniques is typically a set of atmospheric data measured by a collection of monitoring stations spatially distributed in a territory. These stations provide hourly or daily measurements for one or more air pollutants (e.g. NO2, PM10, O3, …), which are usually analysed one at

a time. However, it's increasingly recognized that there exists a dynamic interplay between different chemical air pollutants in the atmosphere. Consequently, a growing body of literature advocates for a paradigm shift towards a "multi-pollutant approach to air quality" [1], acknowledging the intricate relationships between various pollutants and their combined impact on air quality.

For these reasons, the objective of this paper is to briefly discuss modern methods to jointly study a series of air pollutants over a region. We sketch a possible strategy to explore multi-site multi-pollutant atmospheric monitoring data using covariance matrices and exploiting the fact that they represent an importance instance of elements belonging to a Riemannian manifold (instead of classical Euclidean domain) [2]. The particular case study focuses on the analysis of PM10, PM2.5, and NO2 levels in Lombardy during the time period 2021-2022.

The remaining parts of this article are structured as follows. The air pollution and the monitoring station data are described in Section 2, whereas Section 3 briefly introduces the main concepts underlying the analysis of Riemannian manifolds and summarises the future directions for our work.

## 2 Air quality monitoring data

The point-level information regarding the multivariate air pollution measurements was derived from the *Air Quality e-Reporting database*[1] of the European Environmental Agency (EEA) using the corresponding download service. More precisely, the EEA database aggregates more than 12 years of daily and hourly time series measurements for a panel of air pollutants (e.g. NO2, PM10, O3, …) obtained from numerous monitoring stations scattered among the EU member states and several other cooperating countries (which are collectively named EEA38). The individual measurements and the meta-information regarding the monitoring stations involved in the observation process are uploaded into the EEA database by each single entity through a network named EIOnet (Environmental Information and Observation Network). For the Italian case, this process is performed by an official national agency named ISPRA which collects and aggregates data for each region and autonomous province.

Considering the enormous amount of data stored in the complete database, the EEA distributes the raw time series using a binary size-efficient format named `parquet` that can be processed using a multi-language toolbox named *Apache Arrow*. After downloading the raw atmospheric measurements and matching them with the metadata of the monitoring grid, we filtered the observations recorded during the time period 2021-2022 pertaining to those stations that are located in the Lombardy region and jointly record PM10, PM2.5, and NO2.

The complete set of monitoring stations, (a sample of) the raw time series for the chosen pollutants, and the corresponding covariance matrices are reported in Figure 1. As we can see, there are approximately 30 locations spread in different

[1] The database can be browsed at the following link: https://www.eea.europa.eu/en/datahub/datahubitem-view/3b390c9c-f321-490a-b25a-ae93b2ed80c1. Data downloaded on March 2024

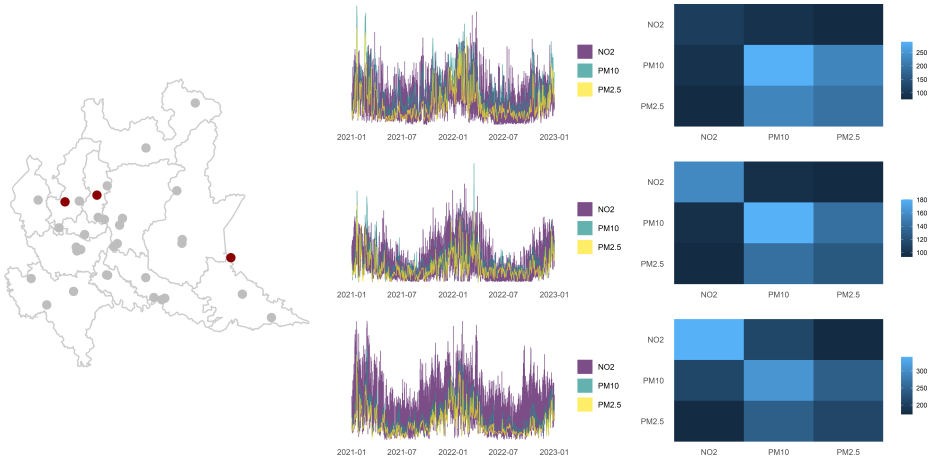


Fig. 1: (Left): Map of the monitoring stations in Lombardy that during the time period 2021-2022 jointly measured PM10, PM2.5, and NO2. Each dot represent a different station. (Center): Time series of the chosen pollutants for the three monitoring stations that are reported as red dots. (Right): Covariance matrices between the pollutants. Each plot corresponds to the adjacent time series.

parts of the region where we jointly observe PM10, PM2.5, and NO2. To simplify the representation, we decided to focus on three of these thirty locations (which are displayed as red dots on the map), graphing the time series of pollutant levels recorded in these stations (central) and the corresponding variance-covariance matrices (right). There are varying degrees of correlation between the three pollutants (especially PM10 and PM2.5).

## 3 Discussion and future works

### Basics of Riemannian Manifolds

The objective of this section is to briefly sketch the main concepts underlying the analysis of data on Riemannian manifolds. In fact, as we already mentioned in the previous sections, symmetric positive-definite matrices (e.g correlation and variance-covariance matrices) are objects lying on a Riemannian manifold, say $\mathcal{M}$, which is a smooth and differential manifold equipped with a *tangent space* at each point $p \in \mathcal{M}$. The tangent space is a vector space consisting of all possible tangent vectors to curves passing through $p$. Its properties let us compute the length of any curve passing through the manifold and, most importantly, find the *geodesic* (i.e. the curve of minimum length) between any two elements of $\mathcal{M}$.

Given a $p \times p$ symmetric positive definite matrix $\mathbf{M} \in \mathcal{M}$, we define its matrix logarithm as

$$\log(\mathbf{M}) = \mathbf{U}\text{diag}\{\log(\lambda_1), \ldots, \log(\lambda_p)\}\mathbf{U}'$$

where $\mathbf{U}$ is the matrix of eigenvectors of $\mathbf{M}$ and $\lambda_1 > \cdots > \lambda_p$ are its eigenvalues. The Frobenius inner product between two matrices $\mathbf{M}_1, \mathbf{M}_2 \in \mathcal{M}$ is defined as

$$< \mathbf{M}_1, \mathbf{M}_2 >_F = \text{tr}(\mathbf{M}_1' \mathbf{M}_2)$$

and the Frobenius norm is equal to $||\mathbf{M}||_F = \sqrt{\text{tr}(\mathbf{M}'\mathbf{M})} = \sqrt{\sum_{i=1}^{n} \lambda_i}$. These constructs let us compute the length of the geodesic between any two symmetric and positive definite covariance matrices:

$$d(\mathbf{M}_1, \mathbf{M}_2) = ||\log(\mathbf{M}_1) - \log(\mathbf{M})_2||_F.$$

The elements introduced in the previous paragraphs are crucial when modelling the dependence between spatial locations by means of (empirical) variogram models for sample covariance matrices. More precisely, given a set $\{s_i\}_{i=1}^n$ of $n$ spatial location in a region $D$ (e.g. the monitoring stations in Lombardy depicted in Figure 1) for which we observed a sample covariance matrix $\mathbf{M}_i \in \mathcal{M}, i = 1, \ldots, n$ (e.g. the covariance matrix among PM10, PM2.5 and NO2), the empirical variogram of the spatial process among the elements in the Riemannian manifold can be computed as

$$\begin{aligned} \hat{\gamma}(h) &= \frac{1}{2|N(h)|} \sum_{(s_i, s_j) \in N(h)} d_g(\mathbf{P}_i, \mathbf{P}_j) \\ &= \frac{1}{2|N(h)|} \sum_{(s_i, s_j) \in N(h)} ||\log(\mathbf{P}_i) - \log(\mathbf{P}_j)||_F \end{aligned} \tag{1}$$

where $N(h) = \{(s_i, s_j) \in D : h - \Delta h < ||s_i - s_j|| < h + \Delta h); i, j = 1, \ldots, n\}$, $\Delta h$ is a positive small quantity, $h > 0$, and $|N(h)|$ counts the number of couples $(s_i, s_j)$ belonging to $N(h)$ [5]. The empirical variogram reported in Equation (1) represent a key step for the estimation of a Kriging model among spatial variance-covariance matrices that respect the underlying geometry of the manifold space.

### Ongoing and future works

Clearly, the techniques described so far represent just the basics for a spatial analysis of air quality data lying on a Riemannian manifold, and we are currently working on testing such methods with the EEA data described in Section 2. Nevertheless, we believe they represent the right approach to tackle the challenging problem of correctly analysing the relationship between a set of air pollutants, taking into account the (geometrically corrected) spatio-temporal dynamics of these processes. In fact, a proper understanding of the (spatio-temporal) relationship and dynamics among pollutants can help decision makers shaping innovative and more effective policies targeting air quality improvement, improve their spatial allocation, and help us better understand the impact of carbon abatement policies.

Finally, we point out that we aim to extend the techniques briefly described in the previous subsection considering, for example, the Kriging framework for

manifold-valued random fields introduced in [5] and, possibly, extending it to a (separable or non-separable) spatio-temporal setting. In this case, we might consider estimating yearly variance-covariance matrices for each monitoring station and adapt the existing methods to the seasonal dimensions depicted in Figure 1. Another interesting line of research raises when studying non-stationary random fields or irregularly shaped spatial domains (e.g. administrative regions with complex boundaries, holes, or barriers). In this case, following the work detailed in [3], we might consider the Random Domain Decomposition approach which will let us overcome the aforementioned problems using an ensemble of local models. Furthermore, the aforementioned technique is able to capture non-stationarities in the residual random field which cannot be explained by only modelling the drift term of the Kriging via a series of covariates (like altitude or population density). This aspect is particularly important when analysing multiple air pollutants since they typically exhibit complex non-linear relationship.

## Acknowledgements

This study was funded by the European Union- NextGenerationEU, in the framework of the GRINS- Growing Resilient, INclusive and Sustainable project (GRINS PE00000018– CUP D43C22003110001). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

## References

1. Dominici, F., Peng, R.D., Barr, C.D. and Bell, M.L., 2010. Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach. Epidemiology, 21(2), pp.187-194.
2. Lee, J.M., 2018. Introduction to Riemannian manifolds (Vol. 2). Cham: Springer.
3. Menafoglio, A., Pigoli, D., Secchi, P., 2021. Kriging riemannian data via random domain decompositions. J. Comput. Graph. Stat. 30 (3), 709–727.
4. Pernigotti, D., Georgieva, E., Thunis, P., and Bessagnet, B. (2012). Impact of meteorological modelling on air quality: Summer and winter episodes in the Po valley (Northern Italy). International Journal of Environment and Pollution, 50(1-4), 111-119.
5. Pigoli, D., Menafoglio, A., Secchi, P., 2016. Kriging prediction for manifold-valued random fields. J. Multivar. Anal. 145, 117–131.
6. Smith, A., Hua, J., de Foy, B., Schauer, J.J. and Zavala, V.M., 2023. Multi-site, multi-pollutant atmospheric data analysis using Riemannian geometry. Science of The Total Environment, 892, p.164064.