PROCEEDINGS

Edited By: Giada Adelfio and Antonino Abbruzzo

GRASPA 2023



GRASPA-SIS BIENNAL CONFERENCE

The Researcher Group for Environmental Statistics of The Italian Statistical Society

TIES EUROPEAN REGIONAL MEETING

The International Environmetrics Society

Palermo, 10-11 July, 2023

Dipartimento di Scienze Economiche Aziendali e Statistiche, Università degli Studi di Palermo

GRASPA

SIS Società Italiana di Statistica

TIES

dSEAS dipartimento scienze economiche aziendali e statistiche

Università degli Studi di Palermo

Proceedings of the GRASPA 2023 Conference
Palermo, 10-11 July 2023
Edited by: Giada Adelfio and Antonino Abbruzzo
–
Palermo: Università degli Studi di Palermo.

ISBN: 979-12-210-3389-2

Questo volume è rilasciato sotto licenza Creative Commons
Attribuzione - Non commerciale - Non opere derivate 4.0



© 2023 The Authors

Sponsored by:

Contents

1 Keynote Sessions 9

1.1 Modeling Extremal Streamflow using Deep Learning Approximations and a Flexible Spatial Process. *R. Majumder, B. J. Reich and B. A. Shaby* 9

1.2 Estimating Covid-19 transmission time using Hawkes point processes. *F. P. Schoenberg* 11

2 Complex data and models 13

2.1 O2S2 for uncertainty quantification in natural background level concentrations. *A. Menafoglio, L. Guadagnini, A. Guadagnini and P. Secchi* 13

2.2 Penalized multivariate hidden semi-Markov models for time series environmental data. *M. Mingione, P. Alaimo Di Loro, F. Lagona and A. Maruotti* 16

2.3 Statistical analysis of complex networks. *N. Pronello, A. Cucco, R. Ignaccolo and L. Ippoliti* 22

3 Statistical Methods and Environmental Studies 27

3.1 ARPA Sicily: collection of environmental data, for an integrated vision of the regional territory and to support the "One Health" approach. *G. Cuffari and A. Conti* 27

3.2 Exploring the effects of temperature on demersal fish communities in the Central Mediterranean Sea using INLA-SPDE modeling approach. *C. Rubino, G. Adelfio, A. Abbruzzo, M. Bosch-Belmar, F. Colloca, M. Di Lorenzo, F. Fiorentino, V. Gancitano and G. Milisenda* 33

3.3 Analytical Methods of the Respiratory Epidemiological Surveys carried out by the National Research Council. *G. Viegi, S. Maio, G. Sarno, S. La Grutta and S. Baldacci* 39

4 Measuring and modelling of environmental 45

4.1 A Bayesian Time Series Model for Reconstructing Hydroclimate from Multiple Proxies. *N. Cahill, J. Croke, M. Campbell, K. Hughes, J. Vitkovsky, J. E. Kilgallen and A. Parnell* 45

4.2 Some empirical results on nearest neighbour pseudo-populations for resampling from spatial populations. *R.M. Di Biase, A. Marcelli, S. Franceschi and L. Fattorini* 52

4.3 Kriging Riemannian data for environmental applications. *A. Menafoglio, D. Pigoli and P. Secchi* 58

5 Spatio-temporal modelling for urban 61

5.1 Stochastic reconstruction of a spatio-temporal Hawkes process with isotropic excitation: an application to road accidents. *P. Alaimo Di Loro and M. Mingione* 61

5.2 ARPALData: retrieving and analyzing air quality and weather data of Lombardy (Italy). *P. Maranzano and A. Algieri* 68

5.3 Mechanistic spatio-temporal modeling of infectious diseases and crime data on urban environments. *J. Mateu and A. Brizi* 74

6 Advances in statistical ecology 81

6.1 Spatial occurrence models using Integrated Nested Laplace Approximation. *S. Martino, J. Belmont Osuna, J. Illian and H. Rue* 81

6.2 Data analysis of photogrammetry-based mapping: the seacucumbers in the Giglio Island as a case study. *G. Mastrantonio, D. Ventura, E. Casoli, A. Rakaj, G. Jona Lasinio and A. Pollice* 83

6.3 A species distribution modelling framework for combining citizen science data from different monitoring schemes. *J. Belmont Osuna, S. Martino, G. Panunzi, and J. Illian* 84

7 Controlling for unmeasured confounding 85

7.1 On the effect of spatial confounding: an approach based on the theory of quadratic forms. *F. Greco and M. Narcisi* 85

7.2 A Flexible Bayesian Time-varying Coefficient Regression Model in Health Applications. *C. Zaccardi, P. Valentini and L. Ippoliti* 92

7.3 Confounder-Dependent Bayesian Mixture Model: Characterizing Heterogeneity of Causal Effects in Air Pollution Epidemiology. *D. Zorzetto, F. J. Bargagli-Stoffi, A. Canale and F. Dominici* 98

8 Posters 99

8.1 Spatio-temporal analysis of the distribution of larval round sardinella in Mediterranean coastal area. *A. Abbruzzo, A. Granata, B. Patti, A. Cuttitta and M. Torri* 100

8.2 A Novel Spatio-Temporal Estimation Method for Occurrences Over Planar and Curved Regions . *B. Begu , S. Panzeri , E. Arnone and L. M. Sangalli* 107

8.3 Addressing Sequestration Bias in Estimating Malaria Parasite Age Distribution: A Bayesian Statistical Model. *E. Bortolato, P. E. Jacob, and J. A. Watson* 113

8.4 Accounting for Misreporting in Spatial Zero-Inflated Poisson Models. *C. Calculli, S. Arima and A. Pollice* 118

8.5 Penalized quantile regression for spatial distributed data. *C. Castiglione, E. Arnone, M. Bernardi, A. Farcomeni, L. M. Sangalli* 124

8.6 Sign-Flip tests for the nonparametric component in Spatial Regression with PDE regularization. *M. Cavazzutti, E. Arnone, F. Ferraccioli, L. Finos, C. Galimberti, L.M. Sangalli* 130

8.7 Environmental exposure indicators and respiratory health in asthmatic children: a case study. *G. Cilluffo, G. Ferrante, S. Fasola , V. Malizia, L. Montalbano, A. Ranzi , C. Badaloni, G. Viegi and S. La Grutta* 135

8.8 Non-parametric density estimation over linear networks. *A. Clemente, E. Arnone, J. Mateu and L. M. Sangalli* 141

8.9 Selecting the Kth nearest-neighbour for clutter removal in spatial point processes through segmented regression models . *N. D'Angelo and G. Adelfio* 146

8.10 Modeling linkage errors in species diversity estimates: an ABC approach. *D. Di Cecco, and A. Tancredi* 152

8.11 The effect of deforestation on infant health: a multilevel mediation analysis. *C. Di Maria* 158

8.12 Geostatistical methods comparison: an application on the concentration of PM2.5 in Lombardy. *A. Fassò, J. Rodeschini and A. Fusta Moro* 164

8.13 Application of the hierarchical variance decomposition approach to an ecological case study with mixed effects. *L. Ferrari, M. Ventrucci and A. Laini* 170

8.14 Variables selection in a P-spline regression model for flood frequency analysis. *A. Gardini* 176

8.15 Assessment of death risk for asbestos cancers using functional regression models among dockworkers exposed to asbestos in northeastern Italy. *P. Girardi, V. Comiati, V. Casotto, C. Gaetan, M.N. Ballarin, E. Merler and U. Fedeli* 182

8.16 Confidence predictive distributions: an application to temperature forecasting in Veneto. *F. Giummolè and V. Mameli* 188

8.17 Particulate matter PM10 assessment by meteorological conditions to monitor extractive activities' impact: a case of study. *E. Iakimova, F. Condino and F. Domma* 193

8.18 Health and environmental risk perception measurement and classification: a case study in Gela (Sicily). *G. Ilardo, A. Pandolfo, V. Malizia, G. Cilluffo, S. Fasola, A. Bonomolo, G. Piva, L. Cori, G. Viegi and S. La Grutta* 199

8.19 Severe storms events' reproduction in the United States of America: evaluation from the marked self-exciting point processes point of view. *F. Lagona and M. Mingione* 205

8.20 A parametric hidden semi-Markov model for toroidal time-series. *G. Lo Galbo, G. Adelfio and M. Chiodi* 211

8.21 Design-based mapping of land use/land cover classes with bootstrap estimation of precision by nearest-neighbour interpolation. *A. Marcelli, R.M. Di Biase, P. Corona, S.V. Stehman and L. Fattorini* 217

8.22 BayesSizeAndShape: a Julia package for Bayesian estimation of Size and Shape data regression models. *G. Mastrantonio and G. Jona Lasinio* 223

8.23 Space-time clustering of seismic events in Chile. *O. Nicolis, L. Delgado, B. Peralta and M. Díaz* 228

8.24 Forecasting Air Pollutants through Artificial Neural Networks. *E. Nissi and L. Saliaj* 234

8.25 Functional principal component analysis for space-time data. *A. Palummo, E. Arnone, L. Formaggia and L. M. Sangalli* 239

8.26 Association between vegetation index and asthma in children and adolescents in North-Central Europe. *A. Pandolfo, G. Ilardo, V. Malizia, G. Squillacioti, S. Fasola, F. Ghelli, G. Ferrante, S. La Grutta, G. Viegi and R. Bono* 245

8.27 A taxonomy for Random Forest in the spatial regression framework. *L. Patelli, M. Cameletti, N. Golini and R. Ignaccolo* 251

8.28 Fish characteristics and microplastic ingestion: a mediation analysis of fish length and trophic level in Western Mediterranean pelagic demersal fish. *M. Sciandra, A. Albano, A. Plaia, C. Di Maria, C. Andolina, S. Vizzini and M.C. Fossi* 257

8.29 UAV plant image Classification Using Combined Machine Learning And Deep Learning Models. *A. Simonetto, G. Tariku and G. Gilioli* 262

8.30 A Bayesian study of temporal changes in seismicity. *E. Varini and R. Rotondi* 268

8.31 On time lag detection between time series sampled by eddy covariance systems. *D. Vitale and D. Papale* 273

Geostatistical methods comparison: an application on the concentration of $PM_{2.5}$ in Lombardy

A. Fassò[1], J. Rodeschini[1,*] and A. Fusta Moro[1]

[1] *University of Bergamo, Via dei Caniana 2, 24127, Bergamo, Italy; alessandro.fasso@unibg.it; jacopo.rodeschini@unibg.it, alessandro.fustamoro@gmail.com*

[*] *J. Rodeschini, jacopo.rodeschini@unibg.it*

Abstract. *Livestock and fertilisers are widely acknowledged to be responsible for approximately 95% of all ammonia emissions, which undergo chemical reactions in the atmosphere to form fine particulate matter, or $PM_{2.5}$. The Lombardy region is characterised by high concentrations of $PM_{2.5}$ and a high livestock density. The main objective of our study is to identify which model better explains the $PM_{2.5}$ concentration in the Lombardy region at the local level. To achieve our objectives, we will utilise multiple modelling strategies, including a Hidden Dynamic Geostatistical Model (HDGM), a Generalized Additive Mixed Model (GAMM), and a Random Forest model with kriging applied to the residual term (RFRK). Furthermore, a comparison between each model will be performed to improve the interpretation of the models and a cross-validation scheme will be used to compare the prediction performance of the three different models.*

Keywords. *Methods comparison; Geostatistical models; Machine learning interpretation; Air quality modelling*

---

1 Introduction

In the Lombardy region, livestock and fertilisers are responsible for up to 95% of all ammonia emissions [1], which undergo chemical reactions in the atmosphere to form fine particulate matter, or $PM_{2.5}$. Due to its stability in the air and limited circulation in the Po Valley, $PM_{2.5}$ often accumulates at high concentrations and adversely affects air quality. This occurs especially during the cold months when the low altitude of the mixing boundary layer combines with high levels of emissions from endothermic-powered vehicles and domestic heating systems. As a result of this high level of pollution, Lombardy has been found to have one of the highest $PM_{2.5}$-related mortality rates among all regions of Europe, with 100-150 premature deaths per 100.000 inhabitants, according to epidemiological studies [5]. The main objective of our study will be to understand which model better explains the $PM_{2.5}$ behaviour in the Lombardy region at the local level. Lombardy is located in the centre of the Po Valley and shares borders with three other large areas, namely Veneto (East), Piedmont (West) and Emilia-Romagna (South). All four areas are characterised by similar production systems, based on industry and agricultural activities, and share the same climatic conditions. As discussed by [10] and [9], $PM_{2.5}$ concentrations respond non-linearly to changes in ammonia and nitrate emissions reductions across the year and with respect to different chemical regimes. The impact of ammonia on the formation of secondary PM is highly dependent on

environmental conditions. While it may not be a significant precursor in some contexts, even a small amount of ammonia can produce a substantial amount of secondary PM in other situations. To handle the non-linear behaviour of the particulate matter and take into account the spatiotemporal correlation we consider different types of models starting from a geostatistical method approach to a machine learning algorithm, passing through a semiparametric model. In particular, to achieve our objectives, we will utilise multiple modelling strategies, including the Hidden Dynamic Geostatistical Model (HDGM) [3], a Generalized Additive Mixed Model (GAMM) [11], and the Random Forest SpatioTemporal Kriging (RFSTK)[2]. The models will be subsequently compared both in their interpretation of regressors and their predictive ability, the latter measured through a Leave-One-Station-Out (LOSO) cross-validation scheme.

2 Dataset

For this work, we use the Agrimonia dataset [6, 7], a comprehensive daily dataset containing air pollutants concentrations [$\mu g/m^3$], livestock densities [$number/km^2$], weather conditions, emissions fluxes and land use characterisation. The Agrimonia dataset is related to the Lombardy region and its surrounding area, calculated as a buffer around administrative boundaries of 0.3 earth degree. The time period covered by the dataset spans from 2016 to 2021, inclusive. The localisations of the data are referred to the localisations of the stations belonging to the air quality monitoring network. In particular, air quality (AQ) data are sampled at 141 ground-level monitoring stations, irregularly located over the augmented Lombardy region. The dataset contains the daily concentrations of the following air pollutants: PM_{10} and $PM_{2.5}$, NO_x, NO_2, SO_2 and NH_3. Table 1 summarises the variables included in the Agrimonia dataset helpful to model the $PM_{2.5}$ in Lombardy. After removing some $PM_{2.5}$ stations due to lack of data, the resulting network includes 49 $PM_{2.5}$ stations.

Dimension	**Variable**
Air quality (AQ)	$PM_{2.5}$(49 stations)
Weather (WE)	Temperature, Precipitation, Wind speed, Relative humidity, Boundary layer height, Surface pressure
Emission (EM)	NH_3 from livestock manure management, NH_3 from agriculture soil, NH_3 from agriculture waste burning, Total NO_x, Total SO_2
Livestock (LI)	Density of pigs, Density of bovines
Land (LA)	Land use, High vegetation index, Low vegetation index

Table 1: Variables retrieved from Agrimonia dataset useful for modelling air quality in the Lombardy region.

3 Preliminary Analysis

In this study, we want to compare different spatiotemporal models on the $PM_{2.5}$ assessment in the Lombardy region. The use of spatiotemporal models is justified by the spatiotemporal variogram [4] shown in Figure 1 which highlights the spatiotemporal correlation structure among $PM_{2.5}$ concentrations. Indeed, moving away in space and time, the observed variance increases according to the distance. The spatial correlation seems to stabilise at ~ 75[km] while the temporal correlation at around 7 days. The temporal correlation is further investigated through the Partial Auto-Correlation Function (PACF) calculated for

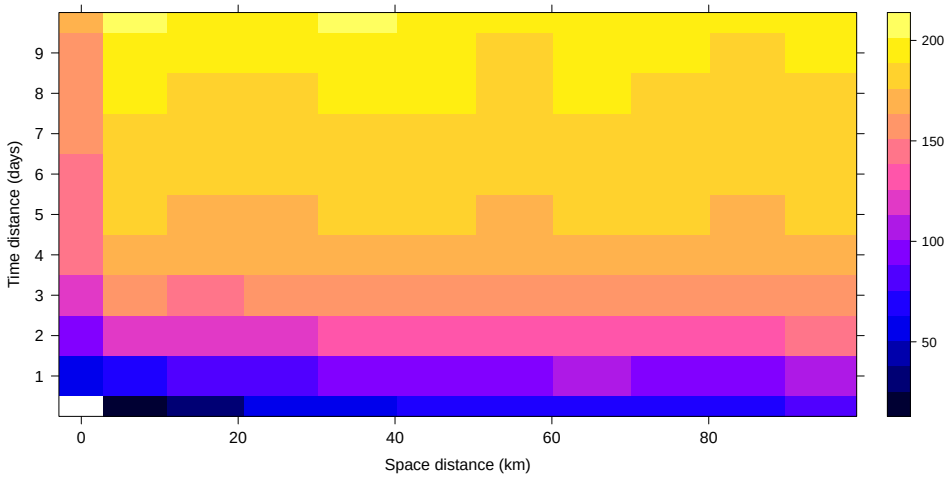


Figure 1: Spatiotemporal variogram [4] computed on $PM_{2.5}$ daily observations at 49 stations from 2016 to 2021. The colour bar represents the sample variance of observations within a certain distance range in space and time.

each station for the first 60 lags (days) and summarised through box plots displayed in Figure 2. The most significant auto-correlated component is lag-1 (1 day). As a result, it is appropriate to use a model with first-order Markovian dynamics. Additionally, the PACF shows that $PM_{2.5}$ is relatively stable in the atmosphere, and the lag-1 is approximately 0.8 for all stations, which indicates that changes in concentration occur smoothly over time. It is worth noting that, in presence of these high levels of correlation both in space and time, a geostatistical model with random effects (as the HDGM) can explain most of the variability even without any regressor. This fact could be problematic for modelling and therefore a challenge is to identify which regressors explain a large amount of the variability of the response variable.

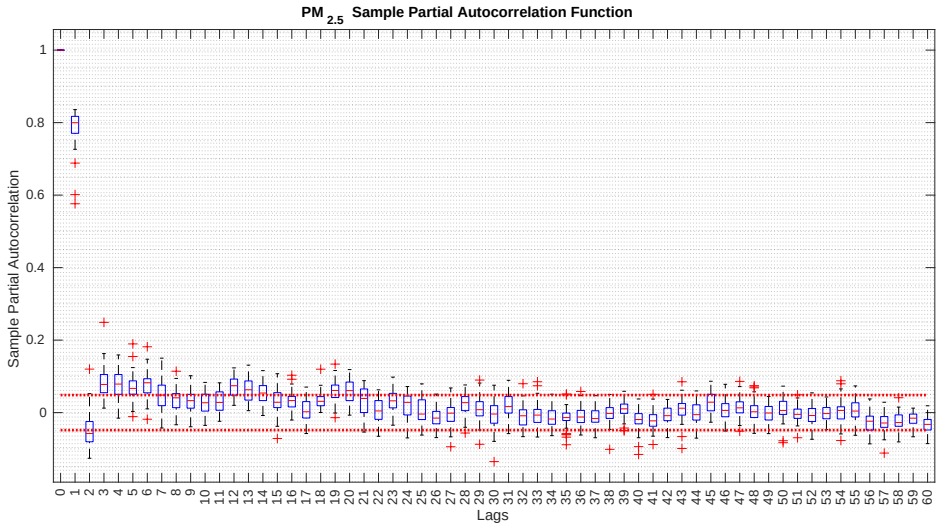


Figure 2: Partial Auto-Correlation Function (PACF) calculated by stations for the first 60 lags (days) and summarised through box plots. The i-th box plot summarises the partial autocorrelation calculated for each station at i-th lag. The lag-1 autocorrelation is approximately 0.8 for all stations, which indicates that changes in concentration occur smoothly over time.

Livestock density data are considered a proxy for emissions from the agricultural sector. However, the air quality monitoring network is not specifically designed to monitor emissions from livestock and therefore many of the monitoring stations are located in areas without any pigs, as shown in Figure 3. So, it will be difficult to estimate the livestock density impact on the $PM_{2.5}$. However, areas presenting high levels of livestock densities are characterised by low levels of air quality (as Cremona and Brescia provinces), highlighting the need for knowledge about the impact of the agricultural sector on air quality.

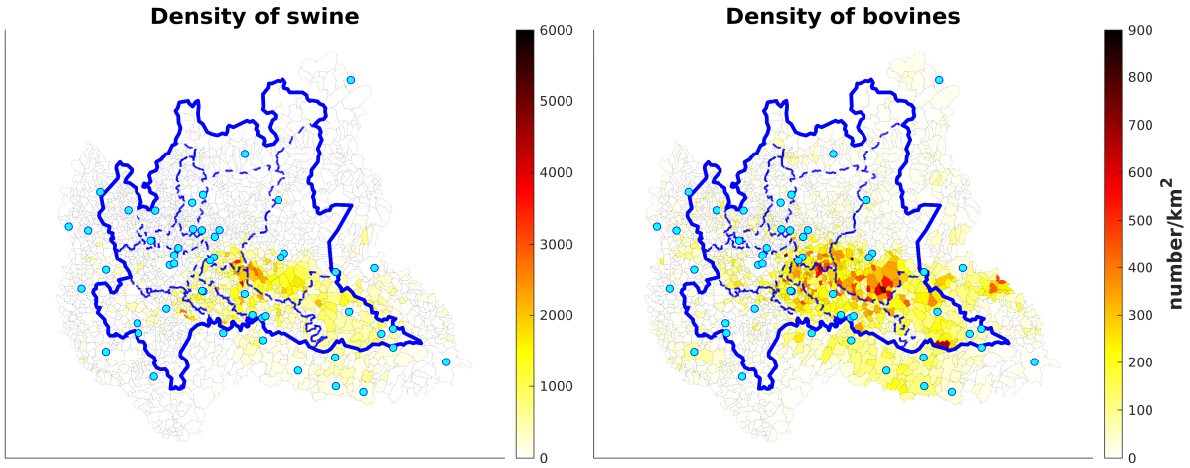


Figure 3: Swine (a) and bovines (b) density over the augmented Lombardy region on 31 December 2021. The AQ stations (cyan circle) are spread randomly over the studied area. The continuous blue line represents the boundary of Lombardy while the dotted blue line represents the Lombardy provinces.

4 Methods

To better understand the relationship between predictor variables and particulate matter, we will consider these types of models: a geostatistical approach, a semiparametric model and a machine learning technique. An important task will be to compare models with each other to understand which model better fits in Lombardy. To achieve our objectives, we will use multiple modelling strategies, including the Hidden Dynamic Geostatistical Model (HDGM) [3], the Generalized Additive Mixed Model (GAMM) [11], and the Random Forest [2] with SpatioTemporal Kriging (RFSTK) [4].

General framework

From a theoretical perspective, consider the $PM_{2.5}$ concentrations as a spatiotemporal stochastic process $\{Z(s,t) : s \in D, t = 1,2,...,T\}$, where D is the spatial domain that contains a set of locations $\{s_i : i = 1,...,n\}$ (ground-level measurement stations) and the temporal domain is $t = 1,...,T$ (daily). Furthermore, we assume that $Z(s,t)$ may be influenced by external variables. So we can express the generic spatiotemporal process as:

$$Z(s,t) = S(s,t) + U(s,t) + \varepsilon(s,t), \tag{1}$$

where $S(s,t)$ are the large-scale components, $U(s,t)$ are small-scale spatiotemporal effects, and $\varepsilon(s,t)$ is the measurement error which is assumed to be a white noise process.

Model description

HDGM is a hierarchical geostatistical model [4] and its usage is demonstrated in the framework

of multivariate air quality modelling [3]. The large-scale consists of linear regression and the small-scale handles the spatiotemporal correlation through a latent random variable with Markovian temporal dynamics and spatial exponential correlation function. The parameters are estimated by the maximum-likelihood (MLE) method using the EM algorithm.

GAMM extends the HDGM by adding smoothing terms on numeric regressors, allowing for non-linear relationships, thus the large scale is a normal GAM. The spatial correlation is modelled within the large scale by adding a smoothing term on the sites coordinates, using a Gaussian process, based on an exponential correlation function. The small scale specifies only the temporal correlation structure, imposing an auto-regressive process of order one.

RFSTK is the combination of a random forest algorithm [2] and spatiotemporal kriging [4]. More precisely, the large-scale effects are obtained by averaging across the estimates of regression trees and the small-scale is treated as a zero-mean Gaussian process with the covariance matrix modelled through a separable space-time exponential correlation function.

The benefit of the geostatistical model vs machine learning:
The geostatistical model is able to estimate the coefficients of the regressor term enabling the inference of the parameters. In this way, it's possible to understand if a covariate is helpful to explain the response variable. In particular, the HDGM and GAMM use the MLE approach to estimate the model parameters and it is possible to compute the associated uncertainty. The GAMM model is more flexible than HDGM, allowing for non-linear regressive effects, estimated as the predictors' smooth functions. The Random Forest (RF) model is a well-known solution to handle the non-linear relationship allowing both non-linear and full interaction regressive terms. However, machine learning techniques are usually referred to as "black-box" because of their lack of interpretability. Furthermore, several approaches have been proposed in the literature to investigate the variable importance, the regression function and the uncertainty associated with the prediction. However, when the number of predictors and the number of tuning hyperparameters (i.e. number of trees, number of random variables chosen for the split, number of observations within final leaves) increase, the computation burden becomes unfeasible.

How to compare the models:
The three different models are comparable in some commonalities but are not entirely comparable due to different definitions. To compare the magnitude of the regressors among models, we compare the estimated coefficients of fixed effects in the HDGM and GAMM to the Variable Importance Factor (VIF) [2] computed within the RF. The VIF is based on the Mean Decrease Accuracy (MDA) algorithm ranking the regressors according to the prediction performance changes after the permutation of the regressor selected. On the other hand, to identify the relationship between the target variable and the regressors, we compare the slope of HDGM fixed effects to the spline of GAMM and the Partial Dependence Plot (PDP) [8] within RF. In particular, the PDP tries to simulate the relationships between predictor variables and predicted responses in the trained dataset. The PDP of predicted responses is computed on a subset of predictor variables X^s by marginalising the other variables X^c (complementary set). Assuming that each observation is equally likely, and the dependence between X^s and X^c and the interactions of X^c and X^s in responses is not strong, the partial dependence can be approximated by using observed predictor data as follows:

$$f^s(X^s) = \frac{1}{N}\sum_{i=1}^{N} f(X^s, X_i^c) \tag{2}$$

where N is the number of observations and $\mathrm{X}_i = (X_i^s, X_i^c)$ is the i-th observation.

Then, we will perform residual diagnostic checking which model best removes the spatial and temporal correlation. In the end, for the assessment of prediction performance, we will use the Leave-One-Station-Out Cross-Validation (LOSO-CV) scheme with a focus on the behaviour of test error and checking for overfitting issues.

5 Further Developments

Using the approaches explained before, we will compare the three different models. Finally, the model selected will be used for scenario analysis, and to estimate the expected effect on PM concentrations under the assumption of a reduction in emissions from agriculture. In particular, the five Shared Socioeconomic Pathways (SSPs), used by the Intergovernmental Panel for Climate Change (IPCC) will be considered along with the Regional Plan of Air Quality Interventions (PRIA). Moreover, cases of efficient manure management, precision agriculture, and a transition to organic farming techniques agreed upon with the stakeholders will be developed.

Acknowledgments. This research was funded by Fondazione Cariplo under the grant 2020-4066 "AgrImOnIA: the impact of agriculture on air quality and the COVID-19 pandemic" from the "Data Science for science and society" program

References

[1] ARPA Lombardia (2022). INEMAR - Inventario Emissioni in Atmosfera: emissioni in Regione Lombardia nell'anno 2019 - versione in revisione pubblica. ARPA Lombardia Settore Monitoraggi Ambientali.

[2] Breiman, L. (2001). Random forests. *Machine learning*, **45**, 5-32.

[3] Calculli, C., et al. (2015). Maximum likelihood estimation of the multivariate hidden dynamic geostatistical model with application to air quality in Apulia, Italy. *Environmetrics*, **26**, 406-417.

[4] Cressie, N., and Wikle, C. K. (2015). Statistics for spatio-temporal data. *John Wiley & Sons*.

[5] European Environment Agency (EEA, 2022). Premature deaths due to exposure to fine particulate matter in Europe (8th EAP). URL: *https://www.eea.europa.eu/ims/health-impacts-of-exposure-to*.

[6] Fassò, A., et al. (2023). Agrimonia: a dataset on livestock, meteorology and air quality in the Lombardy region, Italy. *Scientific Data* **10**, 143.

[7] Fassò, A., et al. (2023). AgrImOnIA: Open Access dataset correlating livestock and air quality in the Lombardy region, Italy - v2.0.2. *Zenodo, URL: https://doi.org/10.5281/zenodo.7593803*.

[8] Friedman, Jerome. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* **29**, 1189-1232.

[9] Nenes, A., et al. (2020). Aerosol pH and liquid water content determine when particulate matter is sensitive to ammonia and nitrate availability. *Atmospheric Chemistry and Physics*, **20**, 3249-3258.

[10] Thunis, P., et al. (2021). Non-linear response of $PM_{2.5}$ to changes in NO_x and NH_3 emissions in the Po basin (Italy): consequences for air quality plans. *Atmospheric Chemistry and Physics* **21**, 9309–9327.

[11] Wood, Simon N. (2017). Generalized additive models: an introduction with R. *CRC press*.