Introduction

In this paper, the authors aim at discussing the role of spatial heterogeneity in spatio-temporal prediction of ground-level air quality (i.e., airborne pollutant concentrations) using both a global and a local approach. The authors refer to a global dataset consisting of all the ground station measurements in Germany and the Netherlands, and to a local dataset comprising only the ground monitoring station in the Amsterdam area. The authors attempt to assess the performance of several algorithms across different spatial scales (global and local) and validate the predictive accuracy when ignoring and when considering local spatial characteristics (i.e., density and population density). The main findings state that that the model performance strongly depends on the considered spatial scale and on the considered spatial locations.

The paper addresses the issue of spatial prediction of air quality in a very broad way and tests several interesting dimensions. However, the work done and the methodology are not rigorous and have several critical points. In particular, several methodological inaccuracies, poor analytical rigor and unclear (if not unwarranted) choices emerged during the reading. Therefore, I suggest that the paper should not be accepted in its present form and should be subject to major revisions (especially in methodology).

General comments:

Hereafter, I state my major concerns that need to be addressed and clarified.

- The overall readability of the text is very poor:
    - the presentation of machine/statistical learning models is very superficial (there are no formulas and no technical modeling aspects are discussed);
    - names and acronyms are inserted into the text without appropriate discussion and description;
    - many sentences need to be rewritten as they are unclear;
    - the paper is very long and confusing: reading requires continuous jumping from one section to another to understand what models and assumptions the authors are analyzing.
- Methodology:
    - Models are presented without specifying their technical characteristics, differences and rationale for their use. No formulas explaining the structure of the models (e.g., the spatiotemporal structure of random effects) are included by the authors. A paper using statistical methodology should never assume that the reader is aware of the methods;
    - Machine learning models (e.g., random forests, xgboost and lightgbm) require great caution and understanding before their use. They are (partially) black-box models with attributes devoted to prediction rather than interpretation of phenomena (while they greatly improve predictions, they also make the results lose interpretive meaning and risk becoming tools that cannot be used by policy makers or practitioners);
    - The expression "linear models" denotes the class of models that are linear in their parameters. It is a very large family. Personally, I struggled to understand which linear models you considered: linear regression? at what scale (original or logarithmic)?

ridge and LASSO are linear, but differ significantly from pure OLS because of the penalty;

- Transformations and distribution of data: the text does not mention the problem of positive skewness (typical of the airborne pollutant concentrations world) (Mudelsee). Clearly, the prediction changes considerably if transformations (e.g., logarithm) are applied to adjust for skewness or if general models with non-Gaussian distributions (e.g., GLMs and GAMs) are used. Where do you stand with respect to this problem?
- There are dozens of linear models and spatio-temporal mixed-effects models in the literature that provide a fair trade-off between interpretability and predictive ability. In the text, none of them are mentioned. I do not intend to cite specific ones, but just type in Google scholar "spatio-temporal models" to retrieve them. I would suggest starting with the spatio-temporal modeling of Wikle-Cressie (who have made history in this branch of research) and colleagues [1, 2];
- The use of cross-validation is relevant for assessing the predictive ability of models. However, remember that random K-fold, as well as LOOCV, are studied for the cross-sectional world. In a spatiotemporal context, they require ad-hoc adjustments that preserve the correlation structure in time and space. In this regard, see the work of Meyer-Pebesma [3-8] on the role of spatial CV and how it is the ideal substitute for random K-fold in the context you study. Also, for the sake of completeness, I suggest adding the word "out-of-sample performances" every time you use CV because it must be clear to reader that all the metrics are computed in a training-test framework to assess predictive capacities of model (and not in-sample fitting);
- Feature selection involves an overwhelming number of alternative techniques: the authors use Shapley values, variables importance, penalty (lasso and ridge), bust subset, etc. Ideally, only one method should be chosen to select relevant covariates so that model results are comparable and not data-dependent. Similarly, it is somewhat problematic to have two or more CV schemes (20-fold and LOO) that prevent proper comparison of the models;
- When reading, I had the feeling (probably wrong) that there was a misunderstanding of the statistical tools used. For example, box plots do not assess the "variance" but the "variability" of a phenomenon and use the median (not the mean) as a reference point, as it is robust to the presence of outliers (frequent in air quality).

Specific comments and technical corrections:

- Abstract, line 2: “model and predict air pollution over space and time”
- Section 2.1, page 5, row 143: the authors state that “We used the precipitation from weather stations (National Centers for Environmental Information, 2017)”. Why not directly using Copernicus ECMWF ERA-5 data, which naturally cover the whole Europe with a fine scale (compared to the study area)? Since you used spatial interpolation (ordinary kriging), how do you account for the interpolation uncertainty generated by this approach? How does it reflect on the following stages?
- Section 2.1, formula after row 165: please, explicitly define the symbols/quantities used in the formula. In its current form, it’s not easy to understand how the traffic is computed;
- Section 2.2

- Row 171: The average median of what? which values are you considering to rank the features? (later on we discover that you take the median of the rankings… but here it is not clear)
- Rows 175-177: Being the first time you cite LASSO, lightgbm and xgboost models, I suggest using the extended names followed by the acronyms. Also, add some theoretical references on the models (e.g., papers explaining the full methodology);
- Row 178: for an extensive comparison in assessing the spatio-temporal prediction accuracy of tree-based methods, linear mixed models and geostatistical mixed models I suggest the papers from the Fassò research group [9-12];

- Section 2.3
  - Actually, the considered models are only described in words but it is difficult to compare it from the analytical perspective. I suggest adding a synoptic table which synthesize the characteristics of the considered models. For instance, the table could state if a model explicitly considers (or not) spatial, temporal or spatio-temporal components (e.g., spatial random effects), if a model is penalized or not (e.g., LASSO), if a model includes covariates or not (e.g., ordinary kriging);
  - Section 2.3.1: can you consider including the recent works on spatial random forests, which extend classical RF to a spatial prediction context [13]? As the aim of the paper is to assess the spatial prediction accuracy of models, this new class could improve a lot your findings;
  - Section 2.3.2, row 214 (α): Please, explicitly define the parameter alpha. Also, if alpha refers to the elastic net mixing parameter, with $0 \leq \alpha \leq 1$, then you are considering the elastic net penalization, which is a combination of LASSO and ridge, and not exactly LASSO or ridge;
- Section 2.4
  - Row 224: "N describes a set of n features" is not clear. which is the difference between N and n? What do you mean by payout? Is it the prediction with the N features? Is S the cardinality of the subset of N?
  - Rows 228-229: please, consider rephrasing the whole sentence: the current sentence seems to state that in general/typically Shapley values are embedded in the two alternative CV approaches. However, this seems to be one of your proposals;
  - Rows 239-242: Please, consider rephrasing the whole sentence as currently it is confused. My interpretation of Figure 2 is that a sensible/remarkable prediction accuracy improvement is obtained when considering at least 12 predictors. However, the improvement is marginal considering more than 12 covariates (the curves become flat);
  - Section 2.4.2 (best subset regression): usually, best subset is used in a linear regression framework. Still, it is not clear to me if you are considering its application in linear or non-linear (in this case, which model?) models. Also, best subset regression is typically affected by computational inefficiency as it requires the computation of 2^k-1 models (where k=30 in the number of covariates). Do you have any insights about the computational burden of this step?
  - Section 2.4.2 (linear models): still, it is not clear to me which class of mixed-effects models are you considering. Please, can you state (in Appendix or Supplementary Materials) the exact formulae and parameter specifications (i.e., which is the structure of the random effects? Are they i.i.d. sequence of Gaussian RVs or are spatio-temporally structured?). Also, later on you state that the "... linear models (i.e., LASSO and ridge) ...": why not considering a multiple linear regression without

penalization? This last model should be directly comparable with penalized approaches.

- Figure 3 (caption): I would say that the upper and lower whiskers provide information about the overall variability of the estimates rather than "variance". Box-whisker plots are typically computed using the IRQ-rule, that is, the whiskers are +-1.5 x IQR (interquartile range, i.e., X_0.75 - X_0.25). Also, box-plots typically use the median as central value. Is the orange line the median? If so, why do you talk about "mean statistic" in the first paragraph of Section 3.1.1? In air quality statistics there is a huge difference among robust (median) and non-robust (mean) methods for assessing the centrality of air quality distributions.
- Section 3.1.1
    - row 286: what do you mean by "uncommon"?
    - right before Table 2: Spatial characteristics is a fundamental feature in air quality statistical modeling. Indeed, local air quality is substantially affected by local weather and environmental conditions. However, why did you not include such variable in the features selection stage? You should be sure about the effective predictive capacity of such variables before including it "a priori". Also, I suppose you used the information through a set of dummy variables (I guess 2 vars). Is that correct? Which one did you choose as reference category? Otherwise, did you used separated/independent models by category (i.e., you estimated all the previous models only for Urban and then for low pop and then for far from road)? In the latter case, you should compare the results with the full dataset very carefully as in the sub-models you are ignoring a large part of the information contained in the full data;
    - row 302: please, clearly state the definition of "more discrete" outcomes or models;
    - rows 307 on: whenever you cite a specific place (e.g., Harleem), please make sure that the area is recognizable on the maps. Where is Harleem in Figures 4 and 5? The same comment holds for all the other cities/locations.
- Figures 4 and 5: they compare different models for different locations. Can you justify this choice? It seems an unfair comparison: to understand the effects of models one should compare different models at the same locations. The comparison you propose is meaningful only if you are sure that, independently on the local conditions, the predictions are comparable (thus there is no spatial effect) and the only relevant factor is the model's definition;
- Figure 6: still, if you use box-plots, then the central value you are comparing is the median. Also, it is not clear to me what you are representing on the box-plots. Are they the distribution of the estimated NO2 concentrations at every point (if so, how many points did you interpolate?) in a specific area (Figures 4 and 5) or are they temporal predictions at some locations (in this case, which locations?) or are the spatio-temporal predictions? Also, where are the results associated with linear mixed models?
- Section 3.1.2: why did you move from a 20-fold CV for global model assessing to a N-fold (LOO) CV for local models? This choice introduces some issues when comparing models as the predictions are computed using sample sizes;
- Table 4: Where are the machine/statistical learning results (i.e., lightgbm, xgboost, random forest)? What is "linear model" and which is its relationship with ridge and LASSO? Why do you compare a different set of models? As for the mapping, if different models are used to compare local and global modeling, the comparison will be biased and unfair;
- Figure 9: are Kerckhoffs's data used to train the models? Why not using the actual NO2 observations (used as response variable of the models) as benchmark? Also, I would plot the

original NO2 concentrations used as response variables in the models. They are the actual benchmark.

Essential bibliography

1. Wikle, C.K., L.M. Berliner, and N. Cressie, *Hierarchical Bayesian space-time models.* Environmental and Ecological Statistics, 1998. **5**(2): p. 117-154.
2. Wikle, C.K., A. Zammit-Mangion, and N. Cressie, *Spatio-temporal Statistics with R*. 2019: Chapman and Hall/CRC.
3. Meyer, H., C. Milà, and M. Ludwig, *CAST: 'caret' Applications for Spatial-Temporal Models*. 2022.
4. Meyer, H., et al., *Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction.* Ecological Modelling, 2019. **411**: p. 108815.
5. Meyer, H., et al., *Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation.* Environmental Modelling & Software, 2018. **101**: p. 1-9.
6. Meyer, H. and E. Pebesma, *Machine learning-based global maps of ecological variables and the challenge of assessing them.* Nature Communications, 2022. **13**(1): p. 2208.
7. Milà, C., et al., *Nearest neighbour distance matching Leave-One-Out Cross-Validation for map validation.* Methods in Ecology and Evolution, 2022. **n/a**(n/a).
8. Meyer, H. and E. Pebesma, *Predicting into unknown space? Estimating the area of applicability of spatial prediction models.* Methods in Ecology and Evolution, 2021. **12**(9): p. 1620-1633.
9. Fassò, A., et al., *Agrimonia: a dataset on livestock, meteorology and air quality in the Lombardy region, Italy.* Scientific Data, 2023. **10**(1): p. 143.
10. Maranzano, P., P. Otto, and A. Fassò, *Adaptive LASSO estimation for functional hidden dynamic geostatistical model.* Stochastic Environmental Research and Risk Assessment, 2023.
11. Maranzano, P. and M. Pelagatti, *Spatiotemporal Event Studies for Environmental Data Under Cross-Sectional Dependence: An Application to Air Quality Assessment in Lombardy.* Journal of Agricultural, Biological and Environmental Statistics, 2023.
12. Otto, P., et al., *Spatiotemporal modelling of PM $ _ {2.5} $ concentrations in Lombardy (Italy)--A comparative study.* arXiv preprint arXiv:2309.07285, 2023.
13. Patelli, L., et al., *A path in regression Random Forest looking for spatial dependence: a taxonomy and a systematic review.* arXiv preprint arXiv:2303.04693, 2023.