https://doi.org/10.5194/egusphere-2024-138
Preprint. Discussion started: 24 January 2024

Random forests with spatial proxies for environmental modelling: opportunities and pitfalls

Carles Milà[1,2], Marvin Ludwig[3], Edzer Pebesma[4], Cathryn Tonne[1,2,5], and Hanna Meyer[3]

[1]Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain
[2]Universitat Pompeu Fabra (UPF), Barcelona, Spain
[3]Institute of Landscape Ecology, University of Münster, Münster, Germany
[4]Institute of Geoinformatics, University of Münster, Münster, Germany
[5]CIBER epidemiología y salud pública (CIBERESP), Madrid, Spain

Correspondence: Carles Milà (carles.mila@isglobal.org)

Abstract. Spatial proxies such as coordinates and Euclidean distance fields are often added as predictors in random forest models; however, their suitability in different predictive conditions has not yet been thoroughly assessed. We investigated 1) the conditions under which spatial proxies are suitable, 2) the reasons for such adequacy, and 3) how proxy suitability can be assessed using cross-validation.

In a simulation and two case studies, we found that adding spatial proxies improved model performance when both residual spatial autocorrelation, and regularly or randomly-distributed training samples, were present. Otherwise, inclusion of proxies was neutral or counterproductive and resulted in feature extrapolation for clustered samples. Random k-fold cross-validation systematically favoured models with spatial proxies even when not appropriate.

As the benefits of spatial proxies are not universal, we recommend using spatial exploratory and validation analyses to determine their suitability, and considering alternative inherently spatial RF-GLS models.

1 Introduction

Predictive modelling of environmental data is key to produce spatially-continuous information out of limited, typically expensive and hard-to-collect point samples. Research fields as diverse as meteorology (Kloog et al., 2017), soil sciences (Poggio et al., 2021), ecology (Ma et al., 2021), and environmental epidemiology (de Hoogh et al., 2018) rely on predictive mapping workflows to produce continuous surfaces, sometimes even at global scale (Ludwig et al., 2023), with products being used for decision-making and subsequent modelling.

Spatial data including environmental variables have intrinsic characteristics that impact the way they are modelled (Longley, 2005). One of the most important is spatial autocorrelation, which as stated by Tobler's first law of geography "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Modellers have used this property to support their spatial interpolation endeavours, which evolved from deterministic univariate approaches such as inverse distance weighting, to more advanced geostatistical methods that can leverage auxiliary predictor information such as Regression Kriging (RK) (Heuvelink and Webster, 2022).

With the increasing availability of spatial information relevant to predict environmental variables (e.g. new satellites and sensors, uncrewed autonomous vehicles, climate and atmospheric simulations), Machine Learning (ML) models have gained momentum in environmental applications due to their ability to capture complex non-linear relationships in highly multivariate datasets (Lary et al., 2016). While standard ML models can better capture complexity in the trend estimation compared to RK, they are *aspatial*, i.e. they ignore the spatial location of the samples and assume independence between observations (Wadoux et al., 2020a). One of the most popular ML algorithms in the geospatial community is Random Forest (RF), a decision tree ensemble (Breiman, 2001) that has shown good performance across many applications (Wylie et al., 2019) and centred the attention of many methodological studies (e.g. Meyer and Pebesma (2021); Hengl et al. (2018); Sekulić et al. (2020); Georganos et al. (2021); Saha et al. (2023)).

The lack of consideration of space in ML models has motivated researchers to find ways to account for spatial autocorrelation to improve model performance. One straightforward approach is to add "spatial proxies" as predictors to the ML model without any modification of the algorithm. We define spatial proxies as a set of spatially-indexed variables with long or infinite autocorrelation ranges that are not causally related to the response variable. The most prevalent type of proxy are coordinates, where either geographical or projected coordinate fields (Fig. 1.3) are added as two additional predictors in the models (e.g. Cracknell and Reading (2014); Walsh et al. (2017); Wang et al. (2017); de Hoogh et al. (2018)). Other approaches include Euclidean Distance Fields (EDF) (Behrens et al., 2018), which, in addition to coordinates, adds five additional EDF with respect to the four corners and centre of the study area (Fig. 1.3); and Random Forest spatial prediction (RFsp) (Hengl et al., 2018), which adds distance fields to each of the sampling locations (Fig. 1.3), i.e. the number of added predictors equals the sample size.

Several advantages of spatial proxy approaches have been discussed by their authors. Hengl et al. (2018) argued that RFsp can address spatial autocorrelation in RF models by accounting for geographical proximity and spatial relations between observations and thus mimick RK. Furthermore, Hengl et al. (2018) pointed out that in RFsp, trend and spatial autocorrelation are modelled in a single step while avoiding the complexities and assumptions of RK. Behrens et al. (2018) explained that with EDF, one can not only account for spatial autocorrelation but also for non-stationarity by means of the partitioning of the geographical space introduced by EDF and its interaction with environmental predictors. To sum up, spatial proxies have been discussed as a straightforward way to address limitations of standard ML methods leading to more accurate predictions.

Nonetheless, the same authors have also expressed caveats. Hengl et al. (2018) warned about using RFsp with clustered data which can result in feature extrapolation, i.e. predicting for values of spatial proxies not included in the training data. Indeed, tree-based models such as RF regression have been acknowledged to perform poorly in feature extrapolation settings (Meyer and Pebesma, 2021; Hengl et al., 2018). Finally, Behrens et al. (2018) showed how RF using coordinates as predictors can result in large artefacts with clearly visible breaks in the predicted surfaces.

Other authors have also expressed views regarding spatial proxies. Meyer et al. (2019) argued that highly autocorrelated variables such as coordinates, especially when used with spatially clustered samples, can result in spatial overfitting leading to poor generalization only detected when using an appropriate spatial Cross-Validation (CV) strategy. Meyer et al. (2019) also showed how spatial proxies typically rank high in variable importance statistics in RF models, especially when they lead to overfitting. Following this, Wadoux et al. (2020a) discussed how high proxy variable importance could hinder correct interpretation of importance statistics for the rest of predictors, which could undermine the possibility to derive hypotheses from the model. Wadoux et al. (2020a) also argued that spatial proxies may hamper residual analysis.

Given this complexity, simulation studies that enable a systematic model evaluation in different controlled environments are needed. To our knowledge, the only simulation study investigating RF with spatial proxies (among other models) is that of Saha et al. (2023), which concluded that RF with coordinates and RFsp performed better than a standard RF. However, their simulation did not reflect the range of characteristics typical of environmental applications as they only explored random sampling designs and did not use spatially-structured predictors. Among other results, Saha et al. (2023) pointed out that methods that add a substantial amount of distance-based predictors such as RFsp will bias the selection of the node-splitting variables toward spatial proxies, leading to poor results when the spatial noise is small compared to the predictor signal.

Even though strengths and weaknesses of spatial proxies have been discussed, a comprehensive assessment of their suitability under different predictive conditions typically found in environmental modelling is still missing. This assessment is important given the broad use of spatial proxies, where coordinates are typically added to the set of predictors by default. We aim to address this knowledge gap by investigating several RF models with spatial proxies, namely coordinates, EDF, and RFsp, with the following objectives:

1. To assess the suitability of spatial proxies in different scenarios regarding the strength of the spatial autocorrelation, the sampling pattern, and predictor availability.
2. To investigate the reasons of such suitability in the different scenarios.
3. To explore whether CV can be used as a model selection tool to guide the choice of spatial proxy.

We address these objectives in a simulation study as well as in two real-world case studies where we modelled air temperature and pollution in Spain.

2 Methods

2.1 Simulation study

We designed a simulation study on a virtual 100x100 square grid to assess, in different prediction settings, the suitability of RF regression models using three different types of spatial proxies: spatial coordinates, EDF, and RFsp (Fig. 1). Our simulation consisted of the following steps:

1. We generated predictor and response surfaces (Fig. 1.1) for different scenarios described in Table 1: partial, where only a subset of the predictors was available; and complete, where all predictors used in the response generation were available for modelling. First, unconditional sequential Gaussian simulation (Gebbers and de Bruin, 2010) was used to generate six independent random fields $\boldsymbol{X}$ with 0 mean and a spherical variogram with sill=1, nugget=0, and range equal to 10

or 40 (see examples in supporting Fig. A1) to be used in response $\boldsymbol{Y}$ generation. Additionally, a noise surface with no spatial autocorrelation was simulated using a standard Gaussian distribution ($\boldsymbol{\mathcal{E}}$, Fig. A1). We generated a response surface using the equation in Table 1.

2. We simulated four sets of training samples (Fig. 1.2) with a sample size of 200 following different distributions: regular samples were drawn by adding random noise (uniform distribution with parameters $U(-2,2)$) to a regular grid, random samples were simulated via uniform random sampling, clustered samples were obtained by simulating 25 (weak clustering) or 10 (strong clustering) randomly-distributed parent points in a first step and 7 (weak) or 19 (strong) offspring points within a 8-unit (weak) or 6-unit (strong) buffer of each parent.

3. For each set of samples, we extracted the corresponding values of the response and predictors, deleted duplicate observations (i.e. two or more points intersecting with the same cell), and fitted a baseline model, which used predictors according to the corresponding scenario (Table 1). We also fitted coordinates, EDF, and RFsp models (see introduction for details) which included the predictors in the baseline model plus the spatial proxies (Fig. 1.3). We fixed the number of trees to 100 and tuned the hyperparameter `mtry` using out-of-bag samples and an equally-spaced grid of length 5 ranging from 2 to the maximum number of predictors.

4. We used each of the fitted models to compute predictions for the entire area and calculated the "true" Root Mean Square Error (RMSE) by comparing the simulated and predicted response surfaces (Fig. 1.4).

5. We also estimated the RMSE using two k-fold CV methods (Fig. 1.5): 5-fold random CV and 5-fold Nearest Neighbour Distance Matching (kNNDM) CV. Briefly, kNNDM is a prediction-oriented method that provides predictive conditions in terms of geographical distances during CV similar to those encountered when using a model to predict a defined area (Linnenbrink et al., 2023; Milà et al., 2022). kNNDM has been shown to provide a better estimate for map accuracy than random k-fold CV when used with clustered samples, while returning fold configurations equivalent to random k-fold CV for regularly and randomly-distributed samples. Estimation of RMSE was done globally to account for the different fold sizes in kNNDM (Linnenbrink et al., 2023), i.e. we stacked all predictions in the different folds and computed the RMSE from all samples simultaneously, rather than computing the RMSE within each fold and then averaging.

6. We computed two additional metrics to understand the feature extrapolation potential and the variable importance of spatial proxies (Fig. 1.6). We calculated the percentage of the study area subject to feature extrapolation as per the Area of Applicability (AOA) (Meyer and Pebesma, 2021) using all training samples. AOA is defined as the area with feature values similar to those of the training data, and is computed based on distances in the predictor space. Unlike feature extrapolation metrics based on variable range or convex hulls, AOA takes into account predictor sparsity within the predictor range and weights variables by their importance in the models. Regarding variable importance, we used the mean decrease impurity method (Breiman, 2002) to quantify the percentage of the total average impurity decrease attributable to spatial proxies.

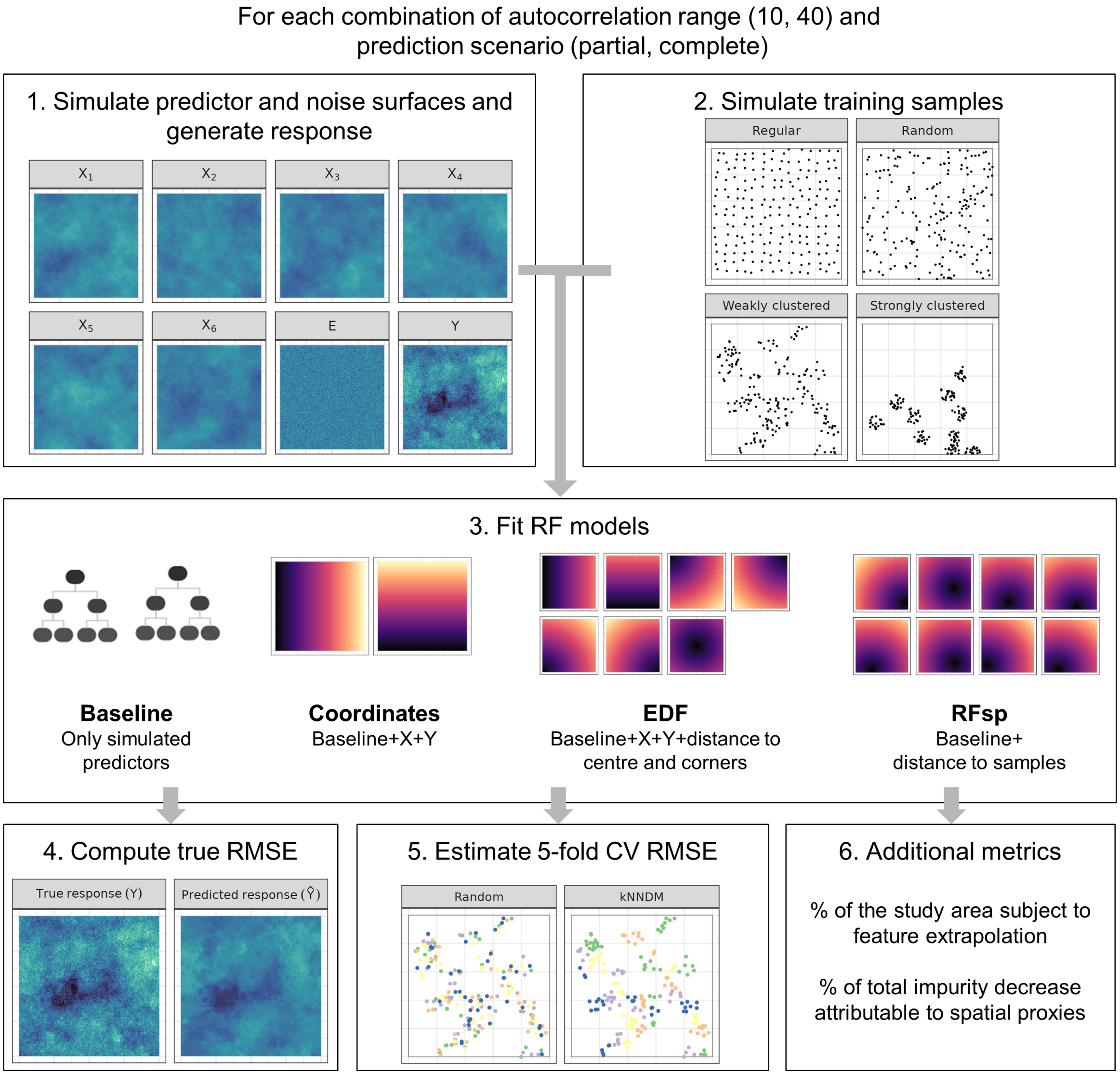

For each combination of autocorrelation range (10, 40) and
prediction scenario (partial, complete)
1. Simulate predictor and noise surfaces and generate response
X1
X2
X3
X4
X5
X6
E
Y
2. Simulate training samples
Regular
Random
Weakly clustered
Strongly clustered
3. Fit RF models
Baseline
Only simulated predictors
Coordinates
Baseline+X+Y
EDF
Baseline+X+Y+distance to centre and corners
RFsp
Baseline+ distance to samples
4. Compute true RMSE
True response (Y)
Predicted response (Ŷ)
5. Estimate 5-fold CV RMSE
Random
kNNDM
6. Additional metrics
% of the study area subject to feature extrapolation
% of total impurity decrease attributable to spatial proxies

Figure 1. Workflow of the simulation study.

120 We ran 100 iterations of each simulation configuration, i.e. we fitted a total of 100 iterations × 2 prediction scenarios × 2 autocorrelation ranges × 4 sample distributions × 4 model types = 6,400 models (without counting the CV fits). We analysed the results of the simulations by plotting the distributions of 1) the true RMSE, 2) the percentage of variable importance attributable to spatial proxies, 3) the percentage of the study area subject to feature extrapolation, and 4) the CV-estimated RMSEs; by each combination of parameters and model type.

Scenario	Description	Response generation equation	Predictors available for modelling
Partial	A subset of predictors $\boldsymbol{X}$ used in the response $\boldsymbol{Y}$ generation are available for modelling	$\boldsymbol{Y} = \boldsymbol{X}_1 + \boldsymbol{X}_2 \cdot \boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 \cdot \boldsymbol{X}_6 + \boldsymbol{\mathcal{E}}$	$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$
Complete	All predictors $\boldsymbol{X}$ used in the response $\boldsymbol{Y}$ generation are available for modelling	$\boldsymbol{Y} = \boldsymbol{X}_1 + \boldsymbol{X}_2 \cdot \boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 \cdot \boldsymbol{X}_6 + \boldsymbol{\mathcal{E}}$	$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4, \boldsymbol{X}_5, \boldsymbol{X}_6$

Table 1. Description of the scenarios considered in the simulation study.

2.2 Case studies

We modelled annual average air temperature and air pollution for continental Spain in 2019 to examine the use of RF models with spatial proxies in real-word examples. For the first case study, we collected daily average air temperature data using the API of the *Agencia Española de Meteorología*, calculated station-based annual averages, and retained 195 stations with a temporal coverage of 75% or higher (Fig. 2). For the second, we collected data on concentrations of Particulate Matter with a diameter of 2.5 microns or less ($PM_{2.5}$) from the *Ministerio para la transición ecológica*. For $PM_{2.5}$ stations with hourly resolution, we first computed daily averages whenever at least 75% of the observations for a given day were available. Then, we computed annual averages and retained 124 stations with an annual temporal coverage of 75% or higher (Fig. 2).

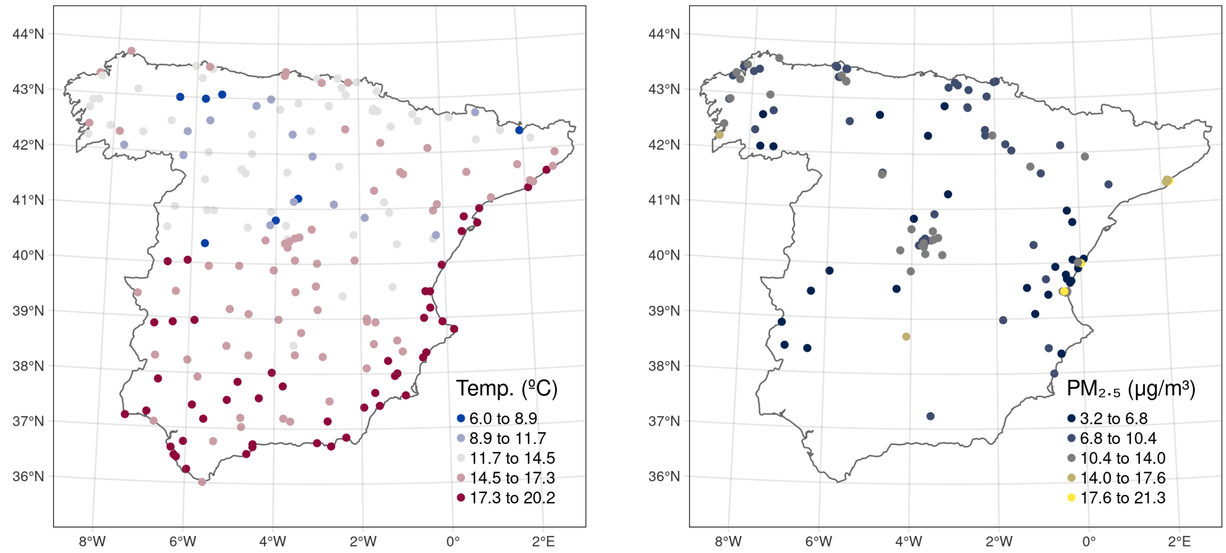


Figure 2. Spatial distribution of the reference station data for the air temperature and air pollution case studies.

We generated a 1 km × 1 km grid covering continental Spain as prediction area. Details of all data used for predictor generation are included in Table A1; while code for all pre-processing steps and processed data used for modelling are publicly available (see code and data availability section below). Briefly, we collected a Digital Elevation Model (DEM), an impervious density product, gridded population counts, land cover data, coastline geometries, road geometries by type, a satellite-based Normalized Difference Vegetation Index (NDVI) from the MODIS Aqua 16-day NDVI product (MYD13A1) and 8-day Land Surface Temperature (LST, MYD11A2) products, annual NightTime Lights (NTL) from VIIRS, and European atmospheric composition reanalyses for $PM_{2.5}$ from Copernicus Atmosphere Monitoring Service (CAMS). We derived population density from the georeferenced population data; we computed % of different land cover classes (urban, industrial, agricultural, natural) in each 1km grid cell; we measured distances from each cell centroid to the nearest coastline; we calculated primary (highway and primary roads) and secondary (all other vehicle roads) road density as the length of the road segments within each 1km cell; we computed annual average composites of the NDVI, LST, and CAMS data. We regridded predictors to the target 1km grid using bilinear interpolation (downscaling) or averaging (upscaling) depending on the source resolution. We extracted predictor values at the station locations for subsequent modelling.

Unlike the simulation study, in these real-world case studies the extent of the spatial autocorrelation of the response and the sample spatial distribution were unknown. To understand how these factors may affect the performance of the different models, we performed an exploratory analysis for each response. First, we assessed the spatial distribution of the monitoring stations using exploratory spatial point pattern analyses. Namely, we estimated the empirical $\hat{G}$, $\hat{F}$, and $\hat{K}$ functions; Monte Carlo simulation (n=99) was used to construct simultaneous envelopes to assess departure from complete spatial randomness (Baddeley et al., 2015). Secondly, we computed empirical variograms of the response variables to assess the strength of the spatial autocorrelation.

For each response, we considered two different sets of variables to be included in the models. First, a naive model, where only one predictor, known a priori to be a strong driver of the response, was used: elevation for temperature and primary road density for $PM_{2.5}$. Second, a complete model, where a much more comprehensive set of predictors was used (see list in supporting Table A1). Our motivation for the naive scenario was to examine whether spatial proxies could help explaining residual spatial autocorrelation due to missing predictors and therefore be used in predictor scarcity settings. Similarly to the simulation study, we used a RF regression baseline model with the selected predictors, as well as coordinates, EDF, and RFsp as additional proxy predictors. We fixed the number of trees to 300 and tuned the parameter `mtry` using out-of-bag samples and an equally-spaced grid of length 10 ranging from 1 to the maximum number of predictors. Using the same methods as in the simulation study, we estimated the performance in terms of global (i.e. calculated in all stacked predictions simultaneously) RMSE and R^2 using 10-fold random and kNNDM CV, calculated the percentage of the study area subject to extrapolation, and estimated the relative importance of spatial proxies. We plotted the predicted surfaces and presented the computed statistics. We assessed residual spatial autocorrelation using empirical variograms of the residuals of each model to evaluate whether spatial dependencies in the data had been captured.

2.3 Comparison with RF-GLS

As an alternative to spatial proxy approaches, we also tested the performance of the Random Forest-Generalized Least Squares (RF-GLS) model recently proposed by Saha et al. (2023). RF-GLS is an extension of RF which relaxes its independence assumption by accounting for spatial dependencies in the data in several ways: 1) they propose a new global dependence split criterion and node representatives to be used during tree construction instead of the CART criterion used in standard RF models; 2) they use contrast resampling rather than bootstrap used in standard RF; 3) they apply residual kriging with covariance modelled using a Gaussian process framework (Saha et al., 2023).

To test the performance of RF-GLS, we included it in the set of candidate models (together with baseline and the three spatial proxy models) in the simulations presented in section 2.1, used it to predict the entire area, and computed the "True" RMSE by comparing the simulated and predicted response surfaces.

2.4 Implementation

Our analyses were carried out in R version 4.2.1 (R Core Team, 2022) using several packages: `sf` (Pebesma, 2018) and `terra` (Hijmans, 2022) for spatial data management; `caret` (Kuhn, 2022), `randomForest` (Liaw and Wiener, 2002), `RandomForestsGLS` (Saha et al., 2022), and `CAST` (Meyer et al., 2023) for spatial modelling; `gstat` (Pebesma, 2004) for random field simulation; and `ggplot2` (Wickham, 2016) and `tmap` (Tennekes, 2018) for graphics and cartographic representations. Additional packages were used for other minor tasks.

3 Results

3.1 Simulation study

Spatial proxies provided little value compared to the corresponding baseline model for the short autocorrelation range, with RMSEs that were similar in the partial scenario with regular, random, and weakly clustered samples; or slightly larger in the complete scenario and for strongly clustered samples (Fig. 3). Nevertheless, for the long range, the addition of spatial proxies resulted in significant reductions in RMSE in the partial scenario except for strongly clustered samples. For the long range and the complete scenario, spatial proxies were irrelevant in terms of performance for regular and random whereas they had a lower performance for clustered samples. Comparing the different types of spatial proxy models, whenever their use was not appropriate, RFsp tended to give worse results than coordinates; nonetheless, together with EDF, it also yielded the largest benefits in partial models with long spatial autocorrelation and regular and random samples.

The relative importance of spatial proxies was most influenced by model type, with spatial proxies having larger importance in models with a higher number of added proxy predictors (Fig. 4). As an example, spatial proxy splits represented a median (IQR) 34% (8.4) of the total impurity decreases for models with coordinates vs. a 84.7% (12.1) for RFsp in the partial scenario with random samples and range=40. Other than that, the relative importance of spatial proxies was greater when samples were strongly clustered, for the long autocorrelation range, and for the partial scenario.

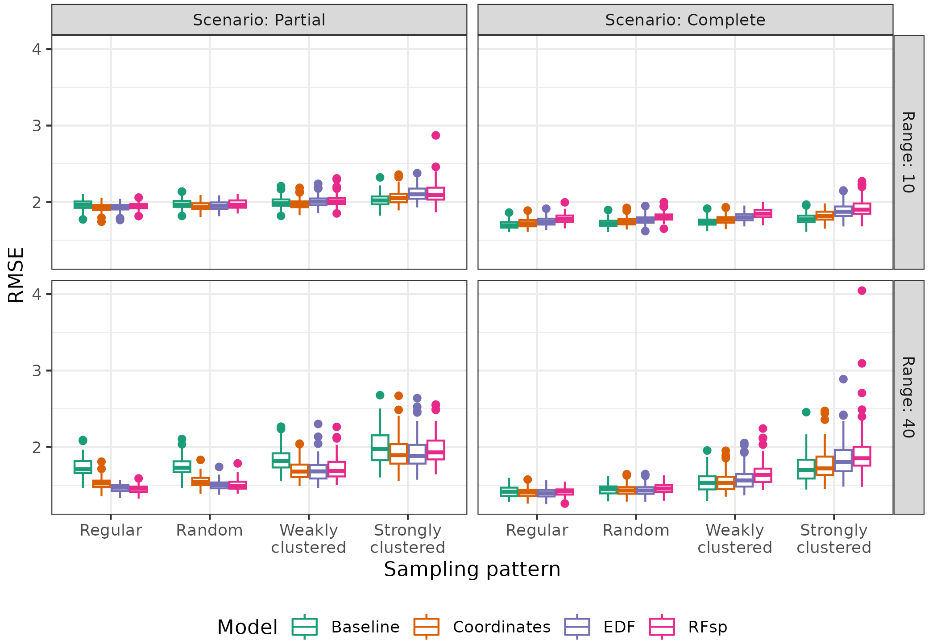


Figure 3. True RMSE (i.e. calculated comparing the entire simulated and predicted surfaces) of each model type by prediction scenario, spatial autocorrelation range, and sampling pattern.

Feature extrapolation was present when samples were clustered whereas it was always low for regular or randomly-distributed samples (Fig. 5). As an example, the median (IQR) percentage of the study area outside the AOA was 60.4 % (15.5) for strongly clustered samples vs. 2.1 % (1.3) for regular samples in partial EDF models with range=10. Within clustered samples, we observed larger feature extrapolation in models with a greater number of spatial proxies (i.e. EDF and RFsp) and a long autocorrelation range.

The two CV methods differed in their ability to estimate the true RMSE as well as in indicating the most suitable model type when samples were clustered. In the partial scenario (Fig. 6) with clustered samples, random 5-fold CV returned underestimated RMSEs and systematically favoured models with spatial proxies although those were not an appropriate choice as indicated by the true RMSE. On the other hand, kNNDM 5-fold CV yielded comparable errors and returned the same model rankings as the ranking based on the true RMSE on median. For regular and random samples, the two CV methods resulted in very similar estimates that were generally well aligned with the true RMSE. Similar findings but with smaller differences were obtained in the complete scenario (supporting Fig. A2).

https://doi.org/10.5194/egusphere-2024-138
Preprint. Discussion started: 24 January 2024

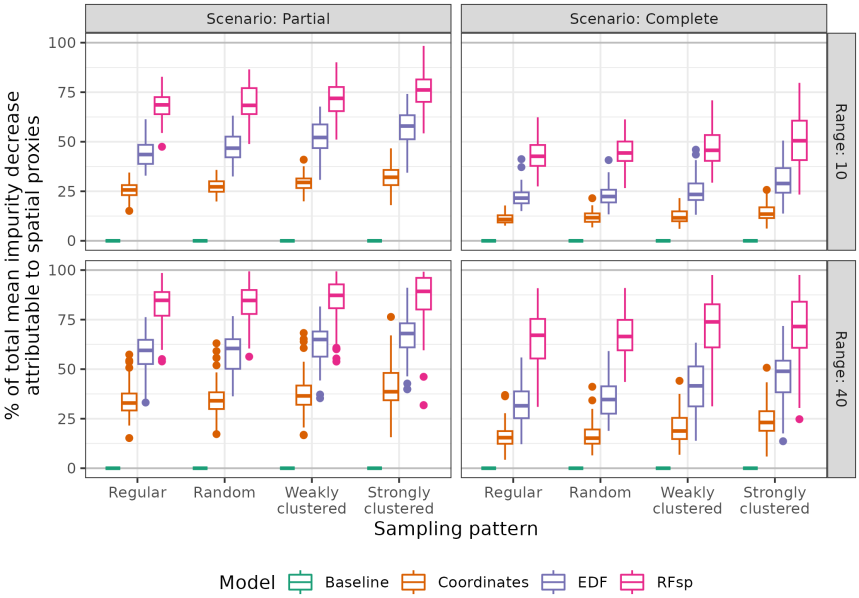


Figure 4. Variable importance of spatial proxies expressed as the percentage of total mean impurity decrease attributable to those variables for each model type by prediction scenario, spatial autocorrelation range, and sampling pattern.

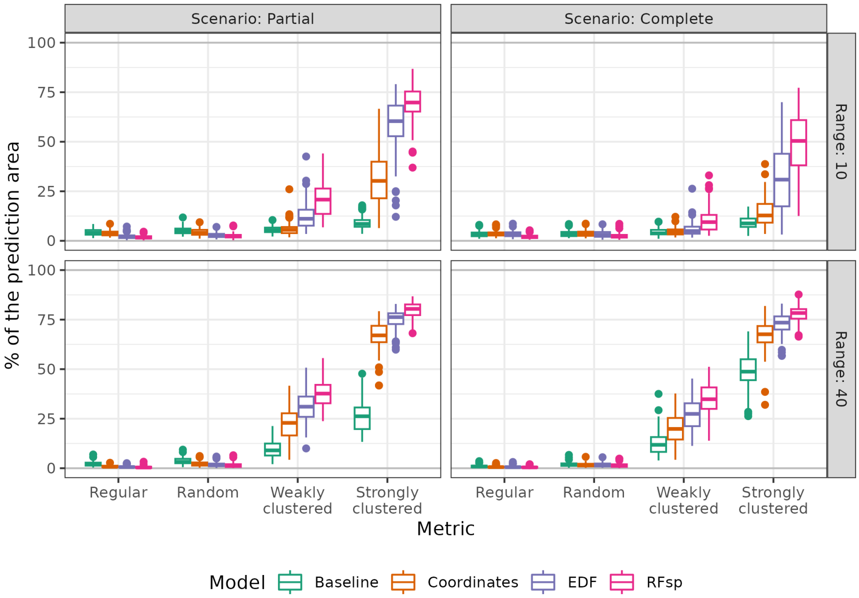


Figure 5. Model feature extrapolation expressed as the percentage of the study area outside of the Area of Applicability (AOA) by prediction scenario, spatial autocorrelation range, and sampling pattern.

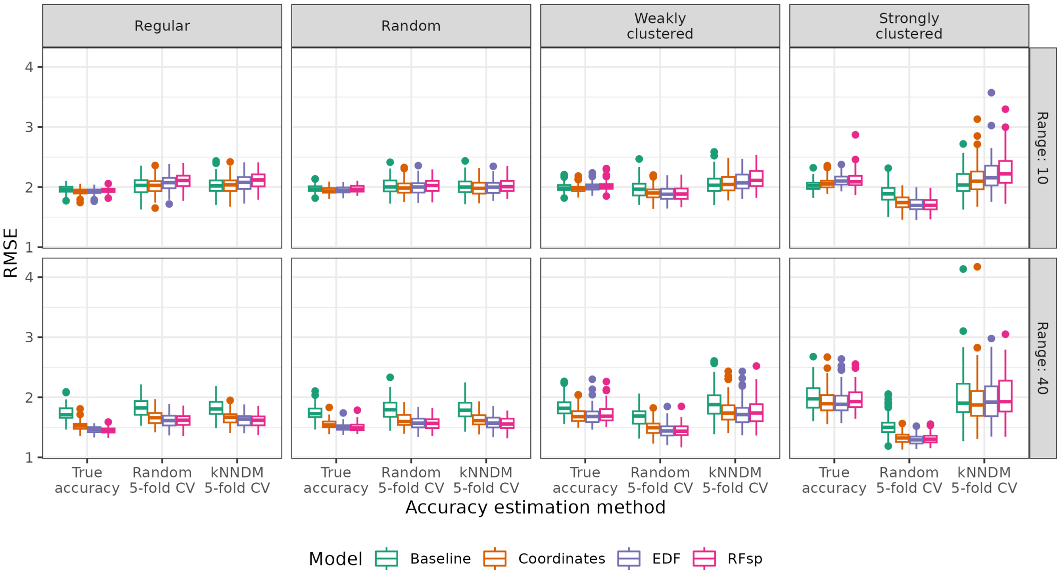


Figure 6. True and estimated (random and kNNDM 5-fold CV) RMSE in the partial prediction scenario by model type, spatial autocorrelation range, and sampling pattern.

3.2 Case studies

Air temperature meteorological stations were well spread within the study area (Fig. 2) and our point pattern exploratory analysis did not suggest a major departure from complete spatial randomness, although there was some evidence of a regular pattern (supporting Fig. A3). Aligned with these results, kNNDM generalised to a random 10-fold CV (supporting Fig. A4).

Results for the naive temperature model indicated substantial gains in performance when using spatial proxies, which yielded only slightly worse results than complete models (Table 2). Performance of all complete models was similar. Feature extrapolation was similar in all cases and lower than 10% of the study area. The importance of spatial proxies was larger in naive models vs. complete models. We detected strong spatial autocorrelation in the response and the residuals of the naive baseline model, which mostly disappeared when adding the whole set of predictors and/or spatial proxies (supporting Fig. A5). Examination of the predicted temperature surfaces indicated that adding spatial proxies to the baseline naive model resulted in similar but somewhat smoother predicted spatial patterns than complete models, while complete models' predictions were almost indistinguishable (Fig. 7).

The distribution of $PM_{2.5}$ stations visually appeared to be spatially clustered (Fig. 2), which was confirmed in our exploratory spatial point pattern analysis with a clear departure from complete spatial randomness (supporting Fig. A6). Reflecting the clustering pattern, the resulting kNNDM had a distinct spatial configuration (supporting Fig. A7).

According to random 10-fold CV, the estimated performance of the baseline naive model in terms of R^2 was null, but it improved substantially when adding spatial proxies. Nonetheless, when using kNNDM CV, the estimated performance was

Model	$\text{RMSE}_{\text{random}}$ (ºC)	R^2_{random}	$\text{RMSE}_{\text{kNNDM}}$ (ºC)	R^2_{kNNDM}	Extrapolation (%)	Proxy importance (%)
Naive						
Baseline	2.04	0.49	2.01	0.51	8.47	0.00
Coordinates	0.97	0.88	0.91	0.90	6.01	50.68
EDF	0.97	0.88	0.92	0.89	7.27	50.53
RFsp	1.07	0.86	1.03	0.87	7.91	65.78
Complete						
Baseline	0.84	0.91	0.81	0.92	6.98	0.00
Coordinates	0.81	0.92	0.80	0.92	6.67	19.33
EDF	0.83	0.92	0.82	0.92	5.44	20.69
RFsp	0.87	0.91	0.87	0.91	4.89	22.29

Table 2. Results of the temperature case study. Subscripts for RMSE and R^2 indicate the type of 10-fold CV used to compute the statistics.

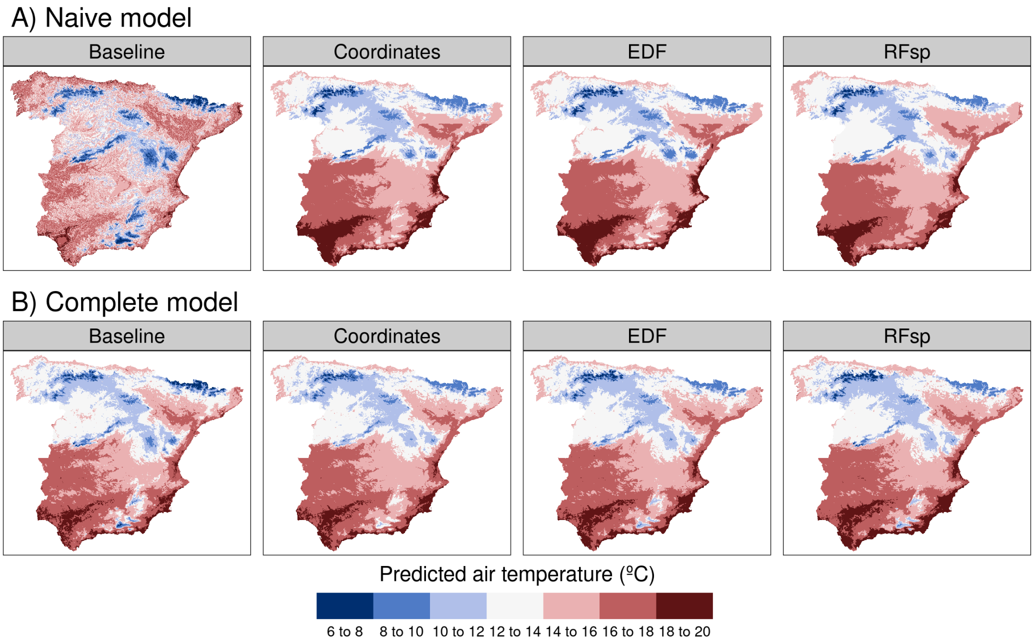


Figure 7. Predicted air temperature using A) naive (DEM only) and B) complete predictors by model type.

similarly low in all cases, thus suggesting significant overfitting (Table 3). Estimated RMSEs of complete models were still lower when using random vs. kNNDM CV, however, the statistics across the different model types were much more similar. Feature extrapolation was the highest in naive models, where proxies had a larger importance that translated into artefacts that were especially evident in the coordinates model (Fig. 8). Predictions for complete models with different spatial proxies

230 were more similar. Inspection of the empirical variograms for the response and residuals of the naive baseline model indicated presence of spatial autocorrelation that was weaker than for air temperature, and which disappeared in complete and spatial proxy models (supporting Fig. A8).

Model	$\mathrm{RMSE}_{\mathrm{random}}$ (μg/m^3)	$\mathrm{R}^2_{\mathrm{random}}$	$\mathrm{RMSE}_{\mathrm{kNNDM}}$ (μg/m^3)	$\mathrm{R}^2_{\mathrm{kNNDM}}$	Extrapolation (%)	Proxy importance (%)
Naive						
Baseline	3.67	0.06	3.78	0.02	1.54	0.00
Coordinates	2.74	0.41	3.66	0.03	11.14	78.45
EDF	2.65	0.45	3.68	0.03	19.76	90.84
RFsp	2.72	0.43	3.92	0.01	19.38	95.02
Complete						
Baseline	2.57	0.48	3.02	0.29	0.60	0.00
Coordinates	2.47	0.52	2.96	0.32	4.64	23.15
EDF	2.50	0.51	3.09	0.27	6.00	38.69
RFsp	2.46	0.53	3.22	0.21	8.65	46.89

Table 3. Results of the $\mathrm{PM}_{2.5}$ case study. Subscripts for RMSE and R^2 indicate the type of 10-fold CV used to compute the statistics.

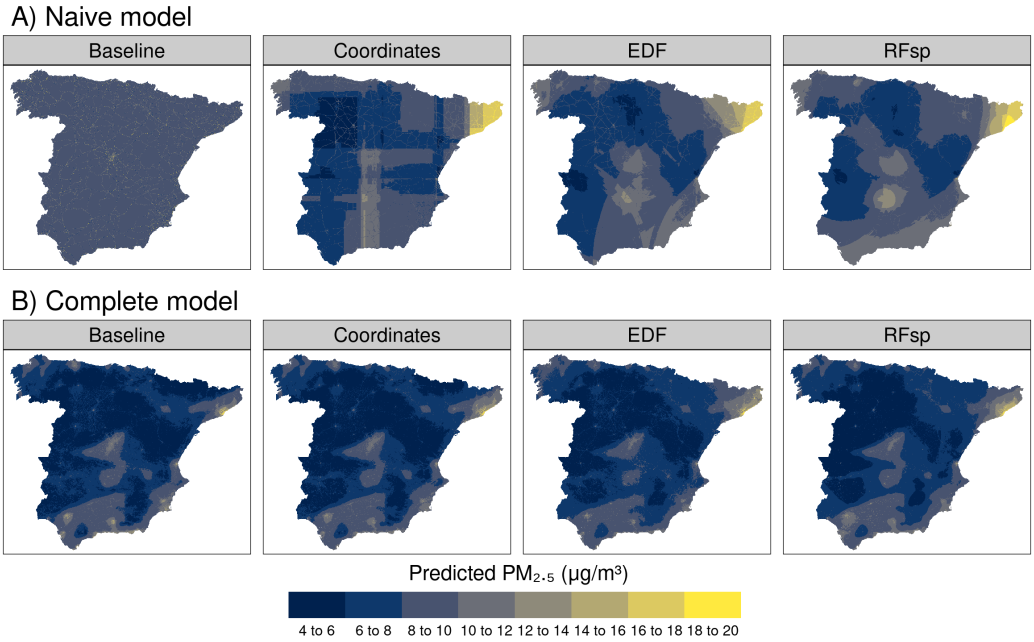


Figure 8. Predicted $\mathrm{PM}_{2.5}$ using A) naive (primary road density only) and B) complete predictors by model type.

3.3 Comparison with RF-GLS

Additional analyses investigating the performance of RF-GLS in the simulation study showed that it outperformed or was on a par with the best-performing standard RF model with and without proxies for all parameter combinations (Fig. 9). While in the complete scenario the performance of RF-GLS was similar to the best-performing standard RF model on median, for the partial scenario RF-GLS was always the superior choice leading to a smaller RMSE.

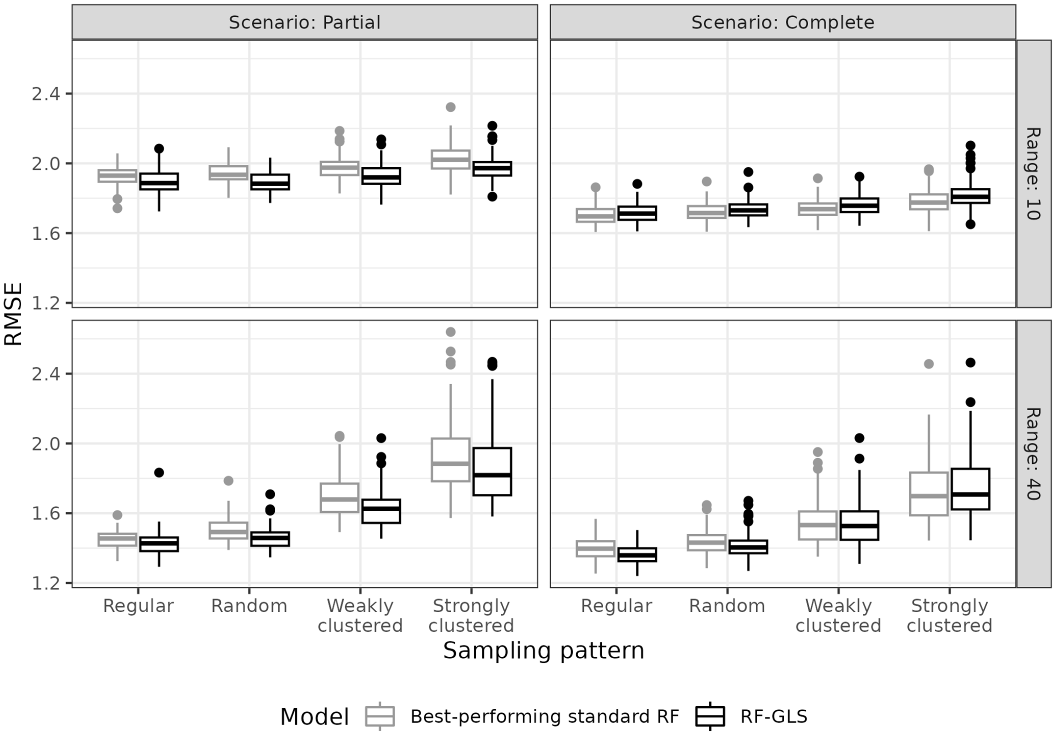


Figure 9. True RMSE (i.e. calculated comparing the entire simulated and predicted surfaces) of the best-performing standard RF for each simulation parameter combination (i.e. the standard RF model with/without proxies with the lowest median RMSE) and RF-GLS, by prediction scenario, spatial autocorrelation range, and sampling pattern.

4 Discussion

Our first two objectives were to identify in which situations RF models with spatial proxies are suitable, and to investigate the reasons behind the observed patterns. Our simulations indicated that an almost necessary condition for proxies to be beneficial is to have regular or randomly-distributed samples over the entire prediction area. This is not surprising since the feature extrapolation potential of spatial proxies with clustered samples has been stressed before (Meyer et al., 2019; Hengl et al., 2018). The more proxies used in the models (e.g. RFsp), the larger the feature extrapolation was. Given these results, although it would be required that spatial proxies had a lower importance when used with clustered samples vs. regular or random, we actually observed the opposite. This is likely a sign of overfitting, where the model uses the proxies to determine the position of the sampling clusters (Meyer et al., 2019), a hypothesis that the difference between the estimated random CV and kNNDM CV supported. Our results are consistent with spatial sampling recommendations for ML models such as RF, which suggest using designs that ensure a good spread in the most important predictors to optimise performance (Wadoux et al., 2019). Hence, spatial proxies are expected to be ill-suited for modelling with clustered samples by design. Even though our simulations indicate that weakly clustered data may sometimes also benefit from spatial proxies in presence of strong residual autocorrelation, we recommend to proceed with caution should this be the case because it may be challenging to define the degree of clustering for which spatial proxies start to be harmful.

Another condition for the proxies to be beneficial is to have a response variable with a long autocorrelation range, reflecting a strong spatial structure. When ranges become shorter, we get closer to the independence assumption of a non-spatial model and thus proxies start to become irrelevant. This is supported by variable importance results, which showed smaller proxy importance for short ranges. Regarding this point, Behrens and Viscarra Rossel (2020) argued that "spatial modeling using a sufficient number predictors (of any kind) with similar or longer ranges (or with coarser scales) than the response variable will produce accurate evaluation statistics, no matter how long the ranges of the predictors are (towards infinity)". Our results suggest that this will be true as long as the spatial structure in the response is strong enough, and samples are randomly or regularly-distributed (see previous paragraph). Experiments for response variables with weaker spatial autocorrelation such as land cover would be interesting follow-up studies to further clarify this point.

Provided that samples are not clustered and response autocorrelation is strong, RF with spatial proxies is beneficial in presence of residual autocorrelation. These potential benefits can be understood by the capacity of spatial proxies to explain residual spatial autocorrelation (Hengl et al., 2018; Behrens et al., 2018), which our results confirmed both in terms of improved performance and removed residual autocorrelation, especially when using a larger number of proxies (EDF or RFsp). However, in the complete scenario where no residual autocorrelation was expected, we hypothesise that the similar or sometimes worse performance is due to adding an irrelevant set of predictors that are noise to the model. Unlike RK, where spatial autocorrelation is modelled in the residuals and in its absence would result in a pure nugget effect, i.e. a flat variogram leading to an ordinary least squares estimation (Hengl, 2007), in a ML model the irrelevant proxies are still included in the trend model. Even though RF is fairly robust to the addition of irrelevant predictors (Kuhn and Johnson, 2019), a decrease in performance was sometimes observed. This idea is supported by the results of Saha et al. (2023), who showed how spatial proxy models such as RFsp perform worse when the spatial noise is small relative to the predictor signal.

Our simulations allow us to give general guidelines on the adequacy of spatial proxies; however, it is important to have a way to confirm them empirically. This was the focus of our third objective, which showed that random CV underestimates map accuracy when used with clustered samples, which has been shown before (Linnenbrink et al., 2023; Wadoux et al., 2021). Perhaps even more important, it incorrectly ranks models, systematically favouring models with a large number of proxies even though those are not always appropriate. On the other hand, kNNDM did provide correct model ranks. We think this is related to overfitting, as, in the presence of clustered sampling, adding spatial proxies may actually help the model to predict at locations geographically close to the samples, as reflected by random CV. Yet, it fails to generalise to the entire prediction area as shown by kNNDM.

Our two case studies had distinct characteristics that impacted the performance of spatial proxy models. Air temperature stations were spread across all the prediction area and measurements exhibited strong spatial autocorrelation. We found that a model with only a DEM and a set of spatial proxies managed to account for the residual spatial autocorrelation and performed almost as well as a much more comprehensive model. This highlights the value of spatial proxies for cost-effective predictive modelling as long as the conditions outlined above are met and the main goal is prediction and not advancing system understanding. For the complete model with a large set of predictors, the inclusion of proxies did neither harm nor benefit the temperature model performance, with predicted surfaces that were very similar.

Regarding air pollution, samples were clustered and the response autocorrelation was weaker. In both naive and complete models, spatial proxies did not improve the performance and large differences in the CV approaches were revealed, highlighting the aforementioned risk of spatial overfitting and wrong conclusions when inappropriate validation practices are used. In the two case studies, we showed the importance of performing a comprehensive spatial exploratory analysis to determine the sample distribution and the response and residual spatial autocorrelation in the baseline model (i.e. without proxies). The results of this analysis can help us determine whether a spatial proxy approach is advisable a priori, which can be confirmed a posteriori using model selection tools such as kNNDM CV.

Our additional analyses regarding the RF-GLS model proposed by Saha et al. (2023) indicated that RF-GLS performed equally or better than the best-performing standard RF with and without spatial proxies in all parameter configurations we considered, while avoiding the complexity of choosing the best set of proxies to be used in each case. We attribute the improved performance of RF-GLS to several reasons; first, in RF-GLS residual variability is modelled as a Gaussian process rather than with spatial proxy predictors in the mean term, which minimizes extrapolation and overfitting problems when spatial proxies are used with clustered samples. Furthermore, in RF-GLS the RF independence assumption is relaxed as spatial autocorrelation is accounted for during the model fitting. Finally, RF-GLS can adapt better to settings where residual spatial autocorrelation is weak or absent since estimation of the covariance function will take the absence of autocorrelation into account, whereas in spatial proxy models all the set of geographical proxies would still be included in the model. All in all, we think that RF-GLS is a step forward in creating truly spatial ML models, and it should be considered as a candidate model in spatial prediction endeavours.

In this study, we included a wide range of predictive scenarios typically encountered in environmental spatial modelling. Nonetheless, there are several points for future work. First, we focused on RF regression and, while we think that our results are likely to extend to other ML algorithms, the extrapolation behaviour and sensitivity to irrelevant predictors differs by algorithm and might limit the ability to generalize our results. Second, our analysis was based on the adequacy of spatial proxies from a prediction accuracy point of view. When using RF for knowledge discovery, variables with long or infinite autocorrelation ranges such as spatial proxies have been identified to be beyond the prediction horizon (Behrens and Viscarra Rossel, 2020; Wadoux et al., 2020b; Fourcade et al., 2018) and variable importance statistics in models including them should be interpreted with extreme caution (Meyer et al., 2019). Third, feature selection based on an appropriate CV scheme has been shown to be helpful to discard irrelevant features prone to overfitting that generalise poorly to new locations such as coordinates (Meyer et al., 2019). In future work, it would be interesting to explore whether feature selection could help to identify irrelevant spatial proxy features in cases where they are not helpful. Finally, the scope of our study was limited to spatial proxies approaches and RF-GLS; however, our analyses could be extended to other models proposed in the literature. Examples include models including spatial lags of the response as prediction features (Sekulić et al., 2020) or geographically-weighted RF (Georganos et al., 2021).

5 Conclusions

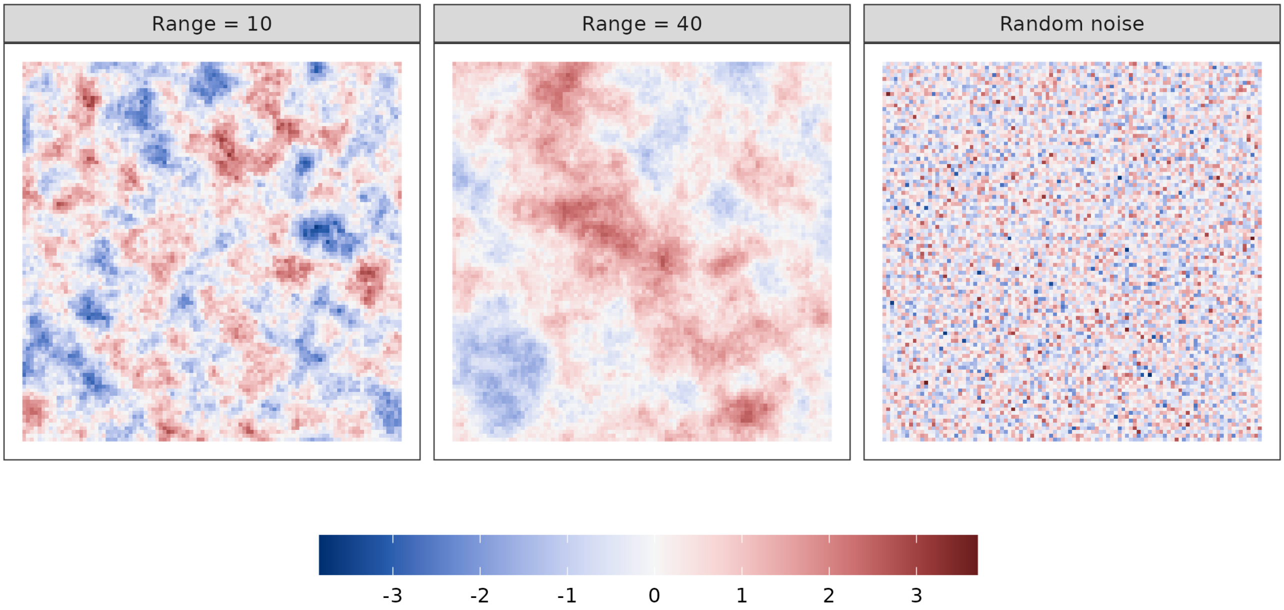
We recommend RF with spatial proxies in cases where both 1) there is presence of residual spatial autocorrelation, 2) samples are regularly or randomly distributed over the study area. In such cases, the addition of spatial proxies is likely to be beneficial in terms or performance. If samples are regular or randomly-distributed but no residual autocorrelation is present, the addition of spatial proxies will have little impact on model performance. Finally, in the presence of clustered samples, using spatial proxies in RF models is not recommended since their inclusion can degrade model performance especially if residual autocorrelation is weak.

More generally, we have shown that the benefits of spatial proxies are not universal and therefore RF modelling with spatial proxies should not be taken as a default approach without careful consideration. Spatial exploratory analysis of the sample distribution and the response and residual autocorrelation are recommended as preliminary steps to evaluate the suitability of spatial proxies, while kNNDM CV can be used as a model selection tool to confirm such suitability by comparing models with and without them, as well as to choose the best set of proxies. Random k-fold CV should not be used for model selection with clustered samples since it systematically favours models with spatial proxies. RF-GLS should be considered as a candidate model for spatial prediction as it performed on a par with or better than standard RF with and without spatial proxies.

Code and data availability. The code for the analysis and the presentation of the results, as well as the data used in the case studies, are available at Milà (2024).

Appendix A: Supplementary figures and tables

Figure A1. Example realizations of random fields used in the simulation study. The first two panels have $\mu = 0$ and a spherical variogram with sill=1, nugget=0, and range indicated in the panel; random noise was generated using a standard Gaussian distribution without spatial autocorrelation.



https://doi.org/10.5194/egusphere-2024-138
Preprint. Discussion started: 24 January 2024

Figure A2. True and estimated (random and kNNDM 5-fold CV) RMSE in the complete prediction scenario by model type, spatial autocorrelation range, and sampling pattern.

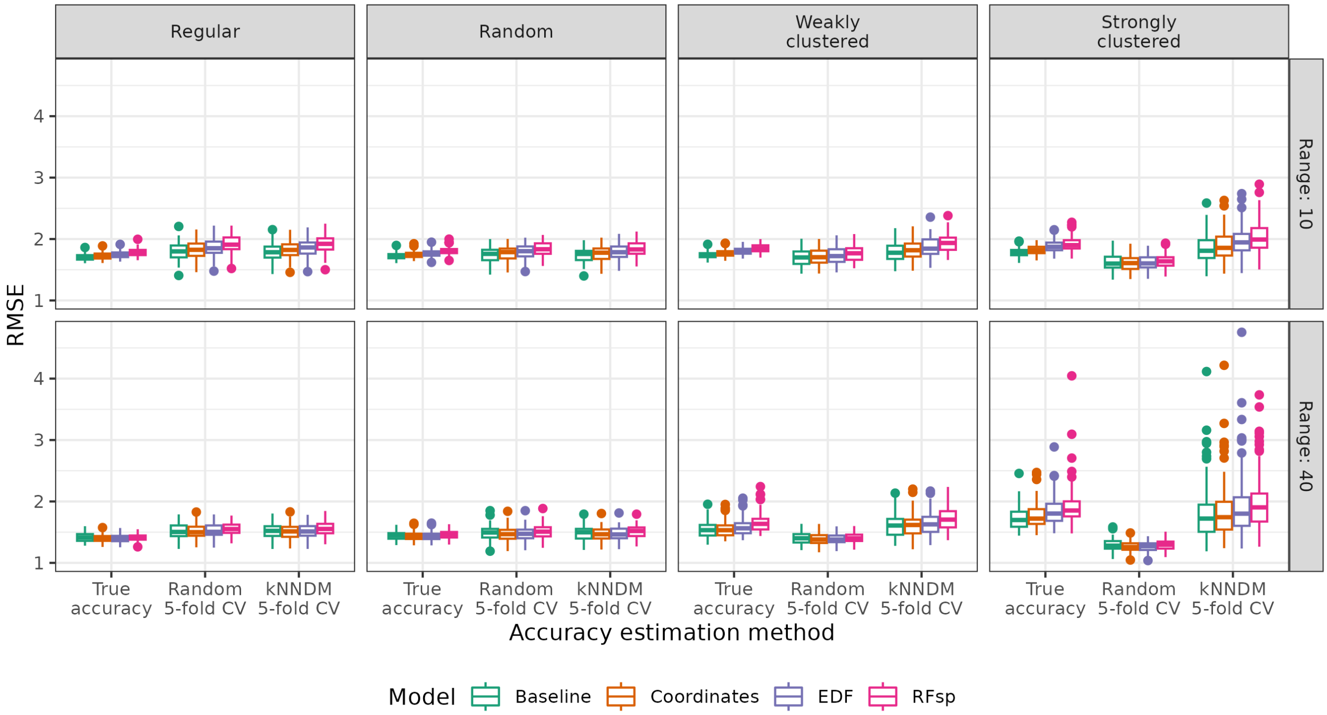
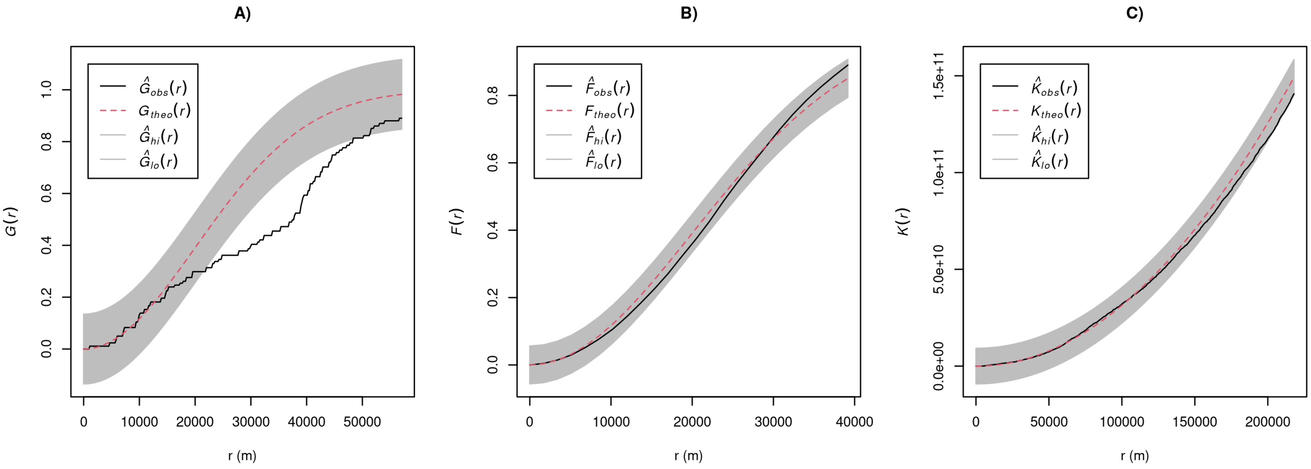
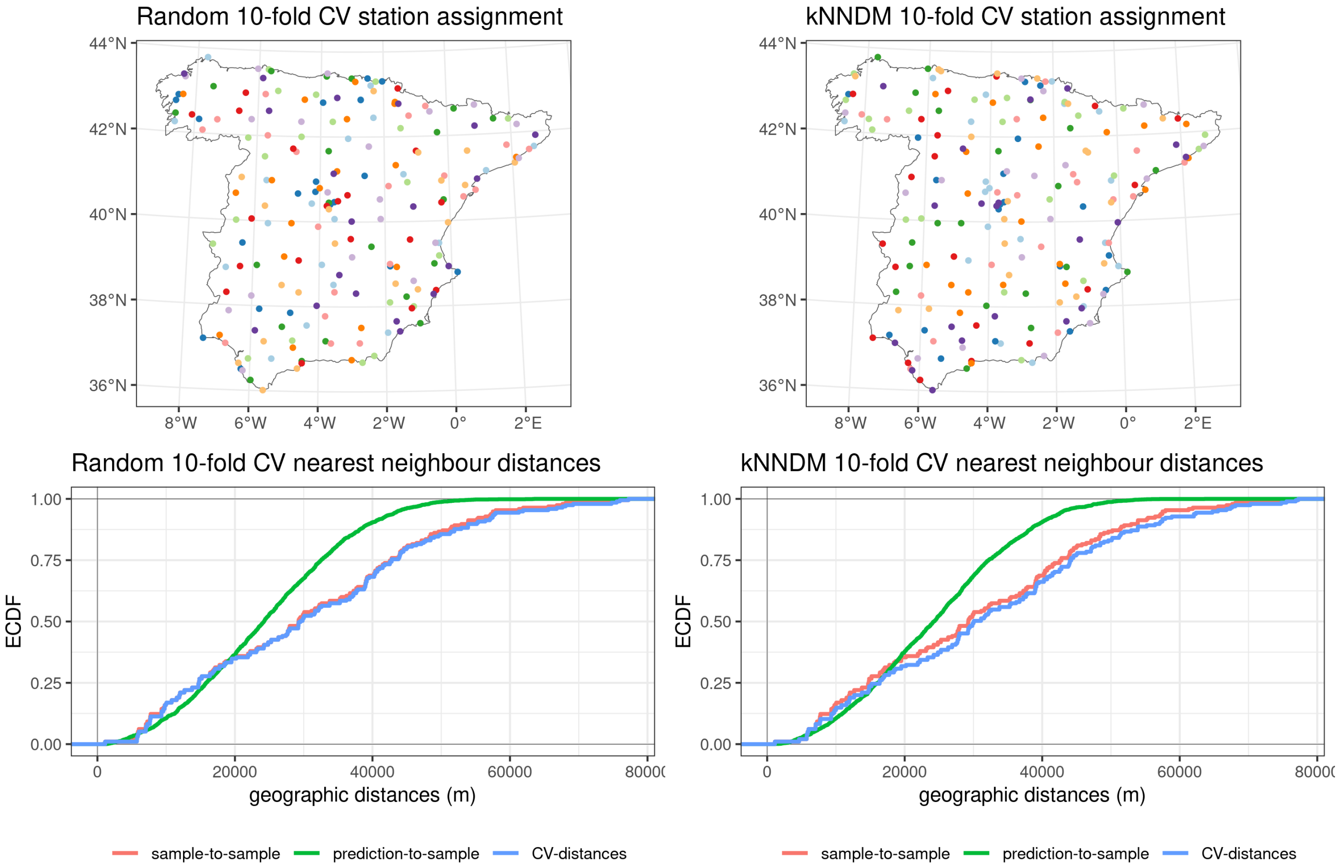


Figure A3. Empirical nearest neighbour distance distribution $\hat{G}$ function (A), empty space $\hat{F}$ function (B), and Ripley's $\hat{K}$ pairwise distance function (C) for the air temperature study case. The dashed red line indicates the theoretical function under complete spatial randomness (i.e. a homogeneous Poisson process) with its global envelope computed using 99 Monte Carlo simulations in grey. Empirical functions calculated from the data are in black.



https://doi.org/10.5194/egusphere-2024-138
Preprint. Discussion started: 24 January 2024

Figure A4. 10-fold assignment according to a random CV method (top left) and the kNNDM method (top right) for the air temperature study case. Figures at the bottom row display the corresponding Empirical Cumulative Distribution Functions (ECDF) of the geographical sample-to-sample, prediction-to-sample, and CV nearest neighbour distances. Ideally, CV-distances should match prediction-to-sample ECDF as much as possible.



https://doi.org/10.5194/egusphere-2024-138
Preprint. Discussion started: 24 January 2024

Figure A5. Empirical variograms for the air temperature response and residuals from all temperature models. Variogram models were fitted for illustrative purposes unless the fit did not converge.

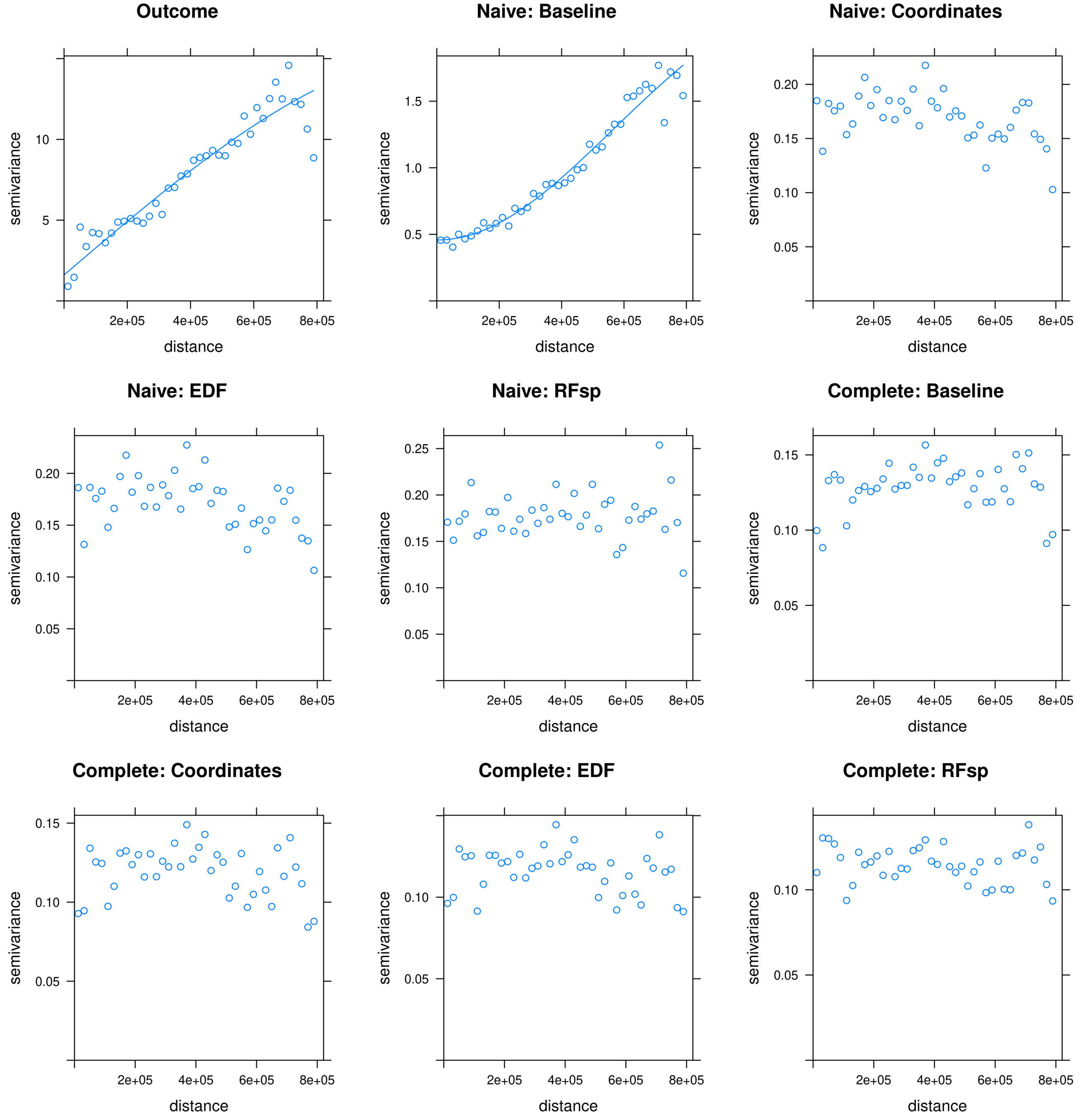
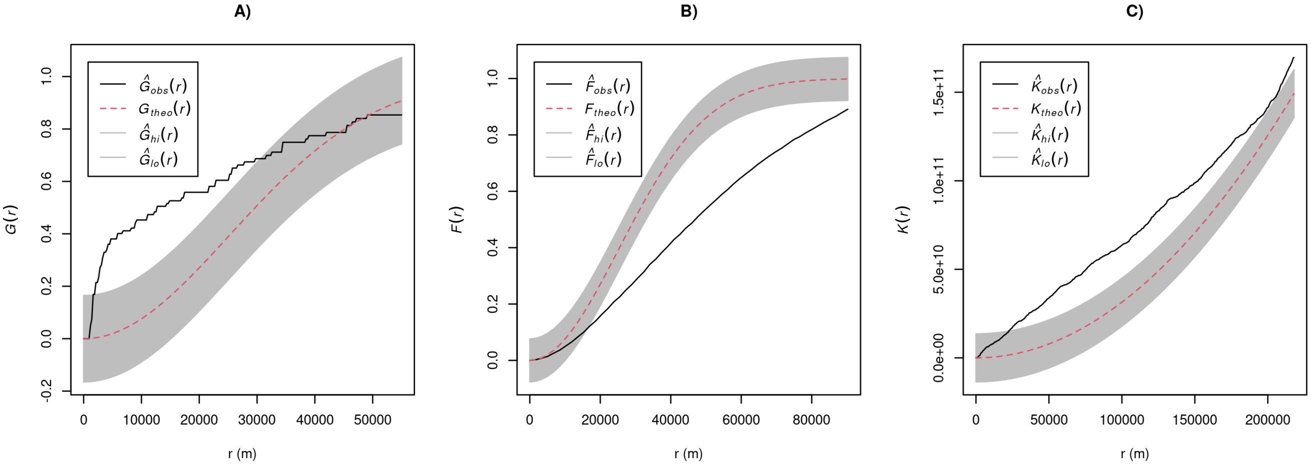
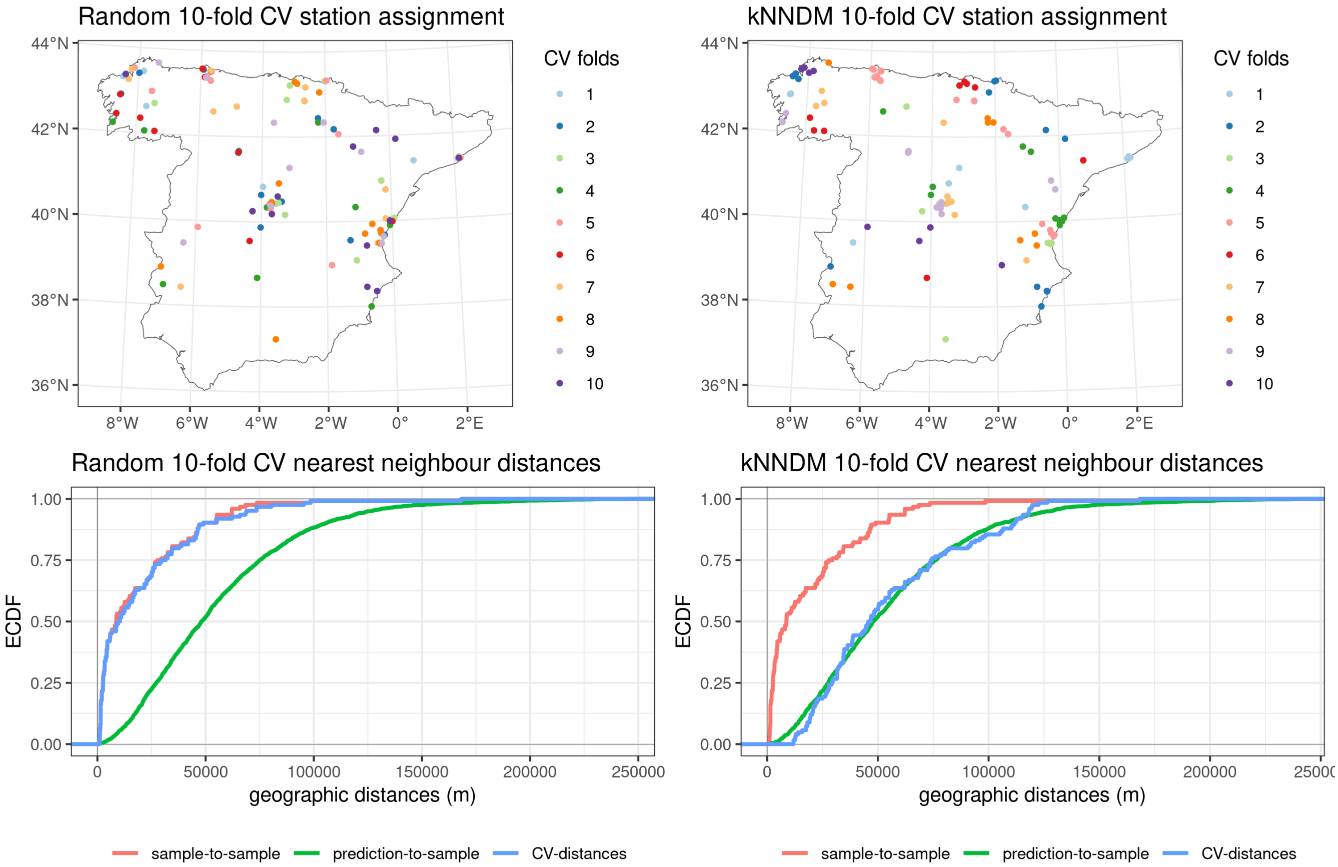


Figure A6. Empirical nearest neighbour distance distribution $\hat{G}$ function (A), empty space $\hat{F}$ function (B), and Ripley's $\hat{K}$ pairwise distance function (C) for the PM$_{2.5}$ study case. The dashed red line indicates the theoretical function under complete spatial randomness (i.e. a homogeneous Poisson process) with its global envelope computed using 99 Monte Carlo simulations in grey. Empirical functions calculated from the data are in black.



https://doi.org/10.5194/egusphere-2024-138
Preprint. Discussion started: 24 January 2024

Figure A7. 10-fold assignment according to a random CV method (top left) and the kNNDM method (top right) for the $PM_{2.5}$ study case. Figures at the bottom row display the corresponding Empirical Cumulative Distribution Functions (ECDF) of the geographical sample-to-sample, prediction-to-sample, and CV nearest neighbour distances. Ideally, CV-distances should match prediction-to-sample ECDF as much as possible.



https://doi.org/10.5194/egusphere-2024-138
Preprint. Discussion started: 24 January 2024

Figure A8. Empirical variograms for the $PM_{2.5}$ response and residuals from all $PM_{2.5}$ models. Variogram models were fitted for illustrative purposes unless the fit did not converge.

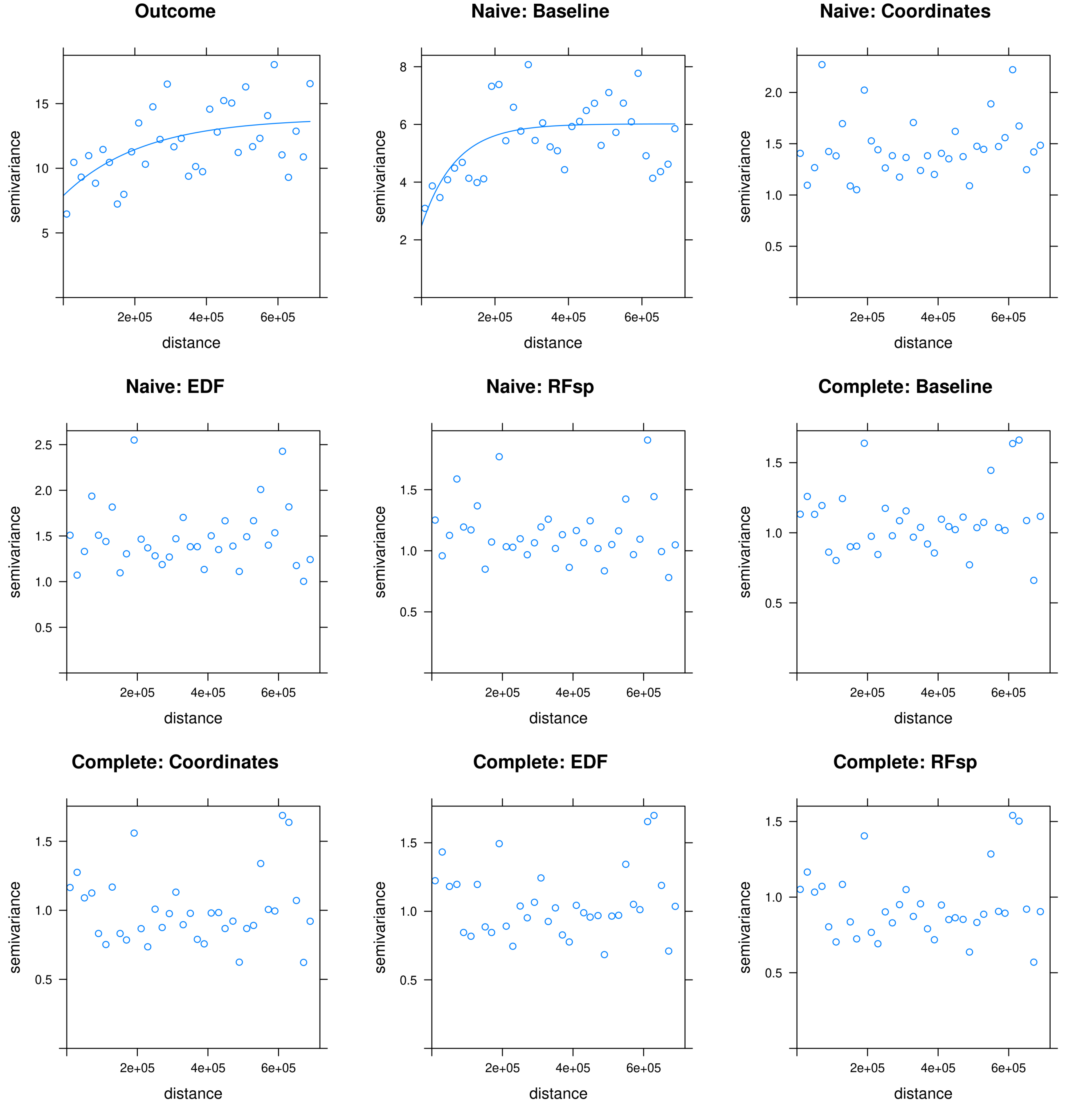


Table A1. List of products and their data source, original spatiotemporal resolution, and use in the complete air temperature and pollution models.

Product	Source	Original resolution	Temperature	$PM_{2.5}$
Station air temperature	Agencia Estatal de Meteorología	Daily	Response	
Station $PM_{2.5}$	Ministerio para la transición ecológica	Hourly/daily		Response
Digital elevation model	CLMS[a]: EU-DEM v1.1	25 m	Predictor	Predictor
Distance to coast	CLMS: EU-HYDRO	Imagery interpretation	Predictor	Predictor
Impervious density	CLMS: IMD (2018)	100 m	Predictor	Predictor
Land Cover	CLMS: CORINE Land Cover (2018)	100 m		Predictor
Population density	Eurostat: GEOSTAT (2018)	1 km		Predictor
Road density	OpenStreetMap	Imagery interpretation		Predictor
NDVI (MYD13A1 v006)	MODIS Aqua Vegetation Indices	16-Day, 500 m	Predictor	Predictor
Nighttime Lights	VIIRS 2019 annual VNL V2 (median)	15 arc second, annual		Predictor
$PM_{2.5}$ reanalysis	CAMS European air quality reanalysis (2019)	0.1°, hourly		Predictor
LST (MYD11A2 v006)	MODIS Aqua Land Surface Temperature	8-Day, 1 km	Predictor	

[a] Copernicus Land Monitoring Service.

Author contributions. All authors participated in the conceptualization and design of the study. CM carried out the analysis, interpreted the results, and wrote the original draft. All authors contributed to discussions, drafts, and gave final approval for publication.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. CM was supported by a PhD fellowship funded by the Spanish Ministerio de Ciencia e Innovación (grant reference: PRE2020-092303). We also acknowledge support from the grant CEX2018-000806-S funded by MCIN/AEI/ 10.13039/501100011033, and from the Generalitat de Catalunya through the CERCA Program.

References

Baddeley, A., Rubak, E., and Turner, R.: Spatial point patterns: methodology and applications with R, CRC press, 2015.

Behrens, T. and Viscarra Rossel, R. A.: On the interpretability of predictors in spatial data science: The information horizon, Scientific Reports, 10, 16 737, 2020.

Behrens, T., Schmidt, K., Viscarra Rossel, R. A., Gries, P., Scholten, T., and MacMillan, R. A.: Spatial modelling with Euclidean distance fields and machine learning, European journal of soil science, 69, 757–770, 2018.

Breiman, L.: Random forests, Machine learning, 45, 5–32, 2001.

Breiman, L.: Manual on setting up, using, and understanding random forests v3. 1, Statistics Department University of California Berkeley, CA, USA, 1, 3–42, 2002.

Cracknell, M. J. and Reading, A. M.: Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information, Computers & Geosciences, 63, 22–33, https://doi.org/https://doi.org/10.1016/j.cageo.2013.10.008, 2014.

de Hoogh, K., Chen, J., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., Bauwelinck, M., van Donkelaar, A., Hvidtfeldt, U. A., Katsouyanni, K., Klompmaker, J., Martin, R. V., Samoli, E., Schwartz, P. E., Stafoggia, M., Bellander, T., Strak, M., Wolf, K., Vienneau, D., Brunekreef, B., and Hoek, G.: Spatial PM2.5, NO2, O3 and BC models for Western Europe – Evaluation of spatiotemporal stability, Environment International, 120, 81–92, https://doi.org/https://doi.org/10.1016/j.envint.2018.07.036, 2018.

Fourcade, Y., Besnard, A. G., and Secondi, J.: Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics, Global Ecology and Biogeography, 27, 245–256, https://doi.org/https://doi.org/10.1111/geb.12684, 2018.

Gebbers, R. and de Bruin, S.: Application of Geostatistical Simulation in Precision Agriculture, pp. 269–303, Springer Netherlands, Dordrecht, ISBN 978-90-481-9133-8, 2010.

Georganos, S., Grippa, T., Gadiaga, A. N., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., and Kalogirou, S.: Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling, Geocarto International, 36, 121–136, https://doi.org/10.1080/10106049.2019.1595177, 2021.

Hengl, T.: A practical guide to geostatistical mapping of environmental variables., Office for Official Publications of the European Communities, https://publications.jrc.ec.europa.eu/repository/handle/JRC38153, 2007.

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., and Gräler, B.: Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables, PeerJ, 6, e5518, 2018.

Heuvelink, G. B. and Webster, R.: Spatial statistics and soil mapping: A blossoming partnership under pressure, Spatial Statistics, 50, 100 639, https://doi.org/https://doi.org/10.1016/j.spasta.2022.100639, special Issue: The Impact of Spatial Statistics, 2022.

Hijmans, R. J.: terra: Spatial Data Analysis, https://CRAN.R-project.org/package=terra, r package version 1.6-47, 2022.

Kloog, I., Nordio, F., Lepeule, J., Padoan, A., Lee, M., Auffray, A., and Schwartz, J.: Modelling spatio-temporally resolved air temperature across the complex geo-climate area of France using satellite-derived land surface temperature data, International Journal of Climatology, 37, 296–304, https://doi.org/https://doi.org/10.1002/joc.4705, 2017.

Kuhn, M.: caret: Classification and Regression Training, https://CRAN.R-project.org/package=caret, r package version 6.0-93, 2022.

Kuhn, M. and Johnson, K.: Feature engineering and selection: A practical approach for predictive models, Chapman and Hall/CRC, 2019.

Lary, D. J., Alavi, A. H., Gandomi, A. H., and Walker, A. L.: Machine learning in geosciences and remote sensing, Geoscience Frontiers, 7, 3–10, https://doi.org/https://doi.org/10.1016/j.gsf.2015.07.003, special Issue: Progress of Machine Learning in Geosciences, 2016.

Liaw, A. and Wiener, M.: Classification and Regression by randomForest, R News, 2, 18–22, https://CRAN.R-project.org/doc/Rnews/, 2002.

Linnenbrink, J., Milà, C., Ludwig, M., and Meyer, H.: kNNDM: k-fold Nearest Neighbour Distance Matching Cross-Validation for map accuracy estimation, EGUsphere, 2023, 1–16, https://doi.org/10.5194/egusphere-2023-1308, 2023.

Longley, P.: Geographic information systems and science, John Wiley & Sons, 2005.

Ludwig, M., Moreno-Martinez, A., Hölzel, N., Pebesma, E., and Meyer, H.: Assessing and improving the transferability of current global spatial prediction models, Global Ecology and Biogeography, 32, 356–368, https://doi.org/https://doi.org/10.1111/geb.13635, 2023.

Ma, H., Mo, L., Crowther, T. W., Maynard, D. S., van den Hoogen, J., Stocker, B. D., Terrer, C., and Zohner, C. M.: The global distribution and environmental drivers of aboveground versus belowground plant biomass, Nature Ecology & Evolution, 5, 1110–1122, 2021.

Meyer, H. and Pebesma, E.: Predicting into unknown space? Estimating the area of applicability of spatial prediction models, Methods in Ecology and Evolution, 12, 1620–1633, 2021.

Meyer, H., Reudenbach, C., Wöllauer, S., and Nauss, T.: Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction, Ecological Modelling, 411, 108 815, https://doi.org/https://doi.org/10.1016/j.ecolmodel.2019.108815, 2019.

Meyer, H., Milà, C., Ludwig, M., and Linnenbrink, J.: CAST: 'caret' Applications for Spatial-Temporal Models, https://github.com/HannaMeyer/CAST, https://hannameyer.github.io/CAST/, 2023.

Milà, C.: Code and data for "Random forests with spatial proxies for environmental modelling: opportunities and pitfalls", https://doi.org/10.5281/zenodo.10495235, 2024.

Milà, C., Mateu, J., Pebesma, E., and Meyer, H.: Nearest neighbour distance matching Leave-One-Out Cross-Validation for map validation, Methods in Ecology and Evolution, 13, 1304–1316, https://doi.org/https://doi.org/10.1111/2041-210X.13851, 2022.

Pebesma, E.: Simple Features for R: Standardized Support for Spatial Vector Data, The R Journal, 10, 439–446, https://doi.org/10.32614/RJ-2018-009, 2018.

Pebesma, E. J.: Multivariable geostatistics in S: the gstat package, Computers & Geosciences, 30, 683 – 691, https://doi.org/https://doi.org/10.1016/j.cageo.2004.03.012, 2004.

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D.: SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty, SOIL, 7, 217–240, https://doi.org/10.5194/soil-7-217-2021, 2021.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/, 2022.

Saha, A., Basu, S., and Datta, A.: RandomForestsGLS: Random Forests for Dependent Data, https://CRAN.R-project.org/package=RandomForestsGLS, r package version 0.1.4, 2022.

Saha, A., Basu, S., and Datta, A.: Random Forests for Spatially Dependent Data, Journal of the American Statistical Association, 118, 665–683, https://doi.org/10.1080/01621459.2021.1950003, 2023.

Sekulić, A., Kilibarda, M., Heuvelink, G. B., Nikolić, M., and Bajat, B.: Random Forest Spatial Interpolation, Remote Sensing, 12, https://doi.org/10.3390/rs12101687, 2020.

Tennekes, M.: tmap: Thematic Maps in R, Journal of Statistical Software, 84, 1–39, https://doi.org/10.18637/jss.v084.i06, 2018.

Tobler, W. R.: A computer movie simulating urban growth in the Detroit region, Economic geography, 46, 234–240, 1970.

Wadoux, A. M.-C., Brus, D. J., and Heuvelink, G. B.: Sampling design optimization for soil mapping with random forest, Geoderma, 355, 113 913, https://doi.org/https://doi.org/10.1016/j.geoderma.2019.113913, 2019.

Wadoux, A. M.-C., Minasny, B., and McBratney, A. B.: Machine learning for digital soil mapping: Applications, challenges and suggested solutions, Earth-Science Reviews, 210, 103 359, https://doi.org/https://doi.org/10.1016/j.earscirev.2020.103359, 2020a.

Wadoux, A. M.-C., Heuvelink, G. B., de Bruin, S., and Brus, D. J.: Spatial cross-validation is not the right way to evaluate map accuracy, Ecological Modelling, 457, 109 692, https://doi.org/https://doi.org/10.1016/j.ecolmodel.2021.109692, 2021.

Wadoux, A. M. J.-C., Samuel-Rosa, A., Poggio, L., and Mulder, V. L.: A note on knowledge discovery and machine learning in digital soil mapping, European Journal of Soil Science, 71, 133–136, https://doi.org/https://doi.org/10.1111/ejss.12909, 2020b.

Walsh, E. S., Kreakie, B. J., Cantwell, M. G., and Nacci, D.: A Random Forest approach to predict the spatial distribution of sediment pollution in an estuarine system, PLOS ONE, 12, 1–18, https://doi.org/10.1371/journal.pone.0179473, 2017.

Wang, Y., Wu, G., Deng, L., Tang, Z., Wang, K., Sun, W., and Shangguan, Z.: Prediction of aboveground grassland biomass on the Loess Plateau, China, using a random forest algorithm, Scientific reports, 7, 6940, 2017.

Wickham, H.: ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York, ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org, 2016.

Wylie, B. K., Pastick, N. J., Picotte, J. J., and Deering, C. A.: Geospatial data mining for digital raster mapping, GIScience & Remote Sensing, 56, 406–429, https://doi.org/10.1080/15481603.2018.1517445, 2019.