Spatial Regression Models for Field Trials: A Comparative Study and New Ideas

Stijn Hawinkel [1,2]*, Sam De Meyer [1,2] and Steven Maere [1,2]*

[1] Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium, [2] VIB Center for Plant Systems Biology, Ghent, Belgium

Naturally occurring variability within a study region harbors valuable information on relationships between biological variables. Yet, spatial patterns within these study areas, e.g., in field trials, violate the assumption of independence of observations, setting particular challenges in terms of hypothesis testing, parameter estimation, feature selection, and model evaluation. We evaluate a number of spatial regression methods in a simulation study, including more realistic spatial effects than employed so far. Based on our results, we recommend generalized least squares (GLS) estimation for experimental as well as for observational setups and demonstrate how it can be incorporated into popular regression models for high-dimensional data such as regularized least squares. This new method is available in the BioConductor R-package *pengls*. Inclusion of a spatial error structure improves parameter estimation and predictive model performance in low-dimensional settings and also improves feature selection in high-dimensional settings by reducing "red-shift": the preferential selection of features with spatial structure. In addition, we argue that the absence of spatial autocorrelation (SAC) in the model residuals should not be taken as a sign of a good fit, since it may result from overfitting the spatial trend. Finally, we confirm our findings in a case study on the prediction of winter wheat yield based on multispectral measurements.

Keywords: prediction, spatial autocorrelation, field trial, generalized least squares, cross-validation, feature selection, simulation

1. INTRODUCTION

Measurements on a cohesive study area, such as performed in field trials (Singh et al., 2003), ecological surveys (Liebhold, 2002), or remote sensing (Wójtowicz et al., 2016), provide a wealth of knowledge on variability occurring between living organisms. On the one hand, the common origin of the data and the use of the same cultivar, soil, weather, and cultivating conditions reduce the variability to manageable levels. On the other hand, there often remains sufficient variability that can be linked to other data types gathered on the same area, such as genomic or optic measurements of the vegetation or environmental features. In observational studies, naturally occurring variability in the outcome variable is linked to naturally occurring variability in the explanatory variables (Bernal-Vasquez et al., 2014; Barmeier and Schmidhalter, 2016; Wójtowicz et al., 2016; Yue et al., 2017; Rocha et al., 2018, 2019; Cruz et al., 2020; Fu et al., 2020; Harkel et al., 2020; Ploton et al., 2020; Tang et al., 2021; Yoosefzadeh-Najafabadi et al., 2021; Zhou et al., 2021). Alternatively, in experimental (interventional) studies, researchers may actively apply a treatment to part of the

study area and investigate the link between this treatment and the variability in the field (Lado et al., 2013; Elias et al., 2018; Mao et al., 2020). For both cases (observational or experimental studies), links between the variables under study can then be tested through formal hypothesis tests and exploited to build, e.g., prediction models for phenotypic traits based on easily measurable variables.

Observations of a given variable across the study area are often not independent but instead exhibit spatial patterns. These patterns may take the shape of small clusters of similar observations, but also of linear gradients or edge effects (despite the use of buffer rows at the border) (Austin and Blackwell, 1980; Langton, 1990; Romani et al., 1993; Haase, 1995; Barmeier and Schmidhalter, 2016; Cruz et al., 2020; Zhou et al., 2021). All of these spatial patterns introduce a dependence between nearby observations, known as spatial autocorrelation (SAC), which violates the assumption of independence among observations of a single variable underlying many statistical methods (Cressie, 1993; Diggle et al., 1998). Confined to a spatial domain, variables that are independent may exhibit an apparent association when they are both spatially autocorrelated, which is known as "spatial confounding" (Hodges and Reich, 2010; Paciorek, 2010; Nobre et al., 2020). As a result, SAC can inflate the type I error rate (Dormann et al., 2007; Kissling and Carl, 2008; Beale et al., 2010), can increase the bias and variability of parameter estimators, and decrease predictive model performance (Dormann et al., 2007; Beale et al., 2010; Rocha et al., 2018, 2019), and can unduly favor the selection of features with spatial structure in feature selection procedures or hypothesis tests (known as "red-shift") (Lennon, 2000; Bini et al., 2009; Meyer et al., 2019; Harisena et al., 2021). Another consequence of SAC concerns data splitting procedures such as cross-validation (CV), whereby the available dataset is repeatedly split into a training and a test set to get an estimate of model performance on new data. The SAC between observations may cause overoptimistic assessment of model performance, as training and test sets are not truly independent (Brenning, 2005; Pohjankukka et al., 2017; Roberts et al., 2017; Meyer et al., 2019; Schratz et al., 2019; Ploton et al., 2020).

A large body of theory is dedicated to spatial interpolation, i.e., leveraging from SAC to improve prediction on new data points lying in between past observations (e.g., kriging) (Cressie, 1993; Diggle et al., 1998). Yet in many contexts (e.g., field trials, ecology), models are trained for application on new datasets gathered in another setup. Spatial coordinates cannot be included in the prediction model in this case, and SAC is a mere nuisance: one wants to learn the relationship between outcome and regressors regardless of spatial effects. Spatial regression models developed for this purpose account for spatial structure in either the mean or error term, but ignore this term when making predictions on new data. A selection of these methods is included and discussed here, but many more exist (see Beale et al., 2010 for a more exhaustive enumeration). Generalized least squares (GLS) is a generalization of ordinary least squares (OLS) to the case of dependent errors, which can be applied for instance to spatial data. Since it requires the estimation of a large number of covariances between observations, additional assumptions are needed to render the estimation feasible. For the case of

spatial data, some functional relationship with a low number of parameters is assumed, which describes the decay of covariance with distance. The corresponding parameters are then estimated from the data (Pinheiro and Bates, 2000). The integrated nested Laplace approximation (INLA) method is based on a similar error model, but performs parameter estimation in a Bayesian context (Selle et al., 2019). Alternatively, neighboring observations can be included as nuisance terms in either the mean or the error model in simultaneous autoregressive models. These models require the definition of a neighborhood based on distance or a certain number of closest neighbors (Kissling and Carl, 2008; Selle et al., 2019). Spatial filtering takes a different approach by performing eigendecomposition on the proximity matrix of the observations and including eigenvectors associated with the outcome variable in the mean model (Tiefelsdorf and Griffith, 2007; Murakami and Griffith, 2015). The spatial structure of the outcome variable can also be modeled more explicitly by including a smooth term in the mean model that is a function of location, e.g., a spline in a generalized additive model (GAM) (Wood, 2003; Rodríguez-Álvarez et al., 2018).

Several simulation studies have investigated how well these methods deal with spatial data. One study to compare a wide range of them was Beale et al. (2010), who found a poor performance for OLS and spatial filtering with forward selection in presence of SAC, and better performance for GAM, GLS, and simultaneous autoregressive models. Kissling and Carl (2008) found simultaneous autoregressive models with spatial terms in the errors to outperform OLS, and Ludwig et al. (2020) found GAM to perform better than OLS in the presence of spatial confounding. Murakami and Griffith (2015) found a better performance of spatial filtering with random effects than with forward selection. The data generation models used in these simulation studies were invariably based on the general assumption of SAC decreasing with the distance between points (Wang and Zhu, 2009; Beale et al., 2010; Alesso et al., 2019; Rocha et al., 2019, 2021; Ludwig et al., 2020; Mao et al., 2020; Harisena et al., 2021). Even when this assumption is acceptable for analysis purposes, this way of drawing observations is not ideally suited to simulate structured spatial patterns observed in some real studies, such as linear gradients or edge effects (Austin and Blackwell, 1980; Romani et al., 1993; Barmeier and Schmidhalter, 2016; Cruz et al., 2020; Zhou et al., 2021).

In field trials, experiments are often designed in anticipation of spatial factors confounding the outcome, e.g., by choosing block designs perpendicular to an expected gradient (Verdooren, 2020). In addition, at the analysis stage, the inclusion of spatial effects in regression models has been shown to improve predictive model performance on real data (Lado et al., 2013; Elias et al., 2018; Mao et al., 2020) and in simulations (Alesso et al., 2019; Rocha et al., 2019; Mao et al., 2020; Harisena et al., 2021), although no improvement was found in another study (Bernal-Vasquez et al., 2014). However, corrections are often done only according to discrete factors of the design, e.g., by adding a correction factor for the rows and columns of checkerboard designs. Yet, the outcome variable may exhibit SAC that is unpredictable at the design stage, within and across subplots. In ecological surveys or when monitoring real cultured

fields, there may even be no fixed design factors at all. Spatial dependence between individual plants or parts of fields then needs to be accounted for in the data analysis rather than in the experimental design. An example is remote sensing of cultured fields through unmanned aerial vehicles (UAVs, colloquially known as drones). The idea is to predict a phenotype such as a yield or biomass based on reflectance measurements. Often vegetation indices are extracted from the raw images as ratios of spectral band intensities to serve as predictors. Although imaging and phenotyping data on a single field are often spatially autocorrelated, this spatial structure is often not taken into account in the data analysis (Yue et al., 2017; Zhang et al., 2019; Fu et al., 2020; Harkel et al., 2020; Lee et al., 2020; Tang et al., 2021; Yoosefzadeh-Najafabadi et al., 2021).

Contemporary genomic or hyperspectral measurements can easily yield many more variables than the number of different plants or plots that were phenotyped. In this high-dimensional context, penalized (or regularized) linear models such as the least absolute shrinkage and selection operator (LASSO) are often employed for feature selection and building predictive models. The question then arises how spatial structure can be incorporated into these models as well. Recently, Cai et al. (2019) combined the LASSO with a moment estimator for autocorrelation in the context of spatial data. Yoon et al. (2013) combined penalized regression with autoregressive modeling. The combination of the GLS framework with the penalized regression methods has also been proposed before, but mostly in other contexts than spatial modeling (Shijun and Benzao, 2006; Yang and Yuan, 2019; Mylona and Goos, 2021). Seya et al. (2015) applied the LASSO to eigenvector selection in eigenvector spatial filtering, and Fan et al. (2017) combined ridge regression with spatial filtering to find that inclusion of spatial effects improves parameter estimation. Yet, Wang and Zhu (2009) found a trade-off between predictive model performance and accuracy of the feature selection, with methods giving better predictions having worse feature selection performance. Software implementations are available for some of these methods (Fan et al., 2017; Yang and Yuan, 2019), but the solutions proposed are usually restricted to a small class of problems.

In view of the multitude of experimental designs, analysis goals, potential spatial patterns, and methods to account for them, deciding on an analysis strategy is not a trivial task. Therefore, we present here a simulation study to assess how various established methods to account for SAC effects in regression models perform (in terms of controlling red-shift, estimating model parameters, selecting relevant features and quantitatively predicting the outcome variable) when confronted with different forms of SAC observed in field or greenhouse trials. We consider experimental studies as well as observational studies and include more realistic spatial structures than in other simulation studies (Dormann et al., 2007; Beale et al., 2010; Murakami and Griffith, 2015; Harisena et al., 2021), such as linear gradients and edge effects. We also combined GLS, which accounts for spatial structure, with penalized regression for high-dimensional data, and investigate how this affects model performance and feature selection. This new method is available as the R-package *pengls* from BioConductor at

https://bioconductor.org/packages/pengls/. Model performance was estimated using random and blocked CV paradigms and compared with performance on independently simulated data. Furthermore, we investigated the role of residual SAC in assessing model fit. To illustrate our findings, we apply spatial and non-spatial methods to a real dataset from remote sensing of wheat fields.
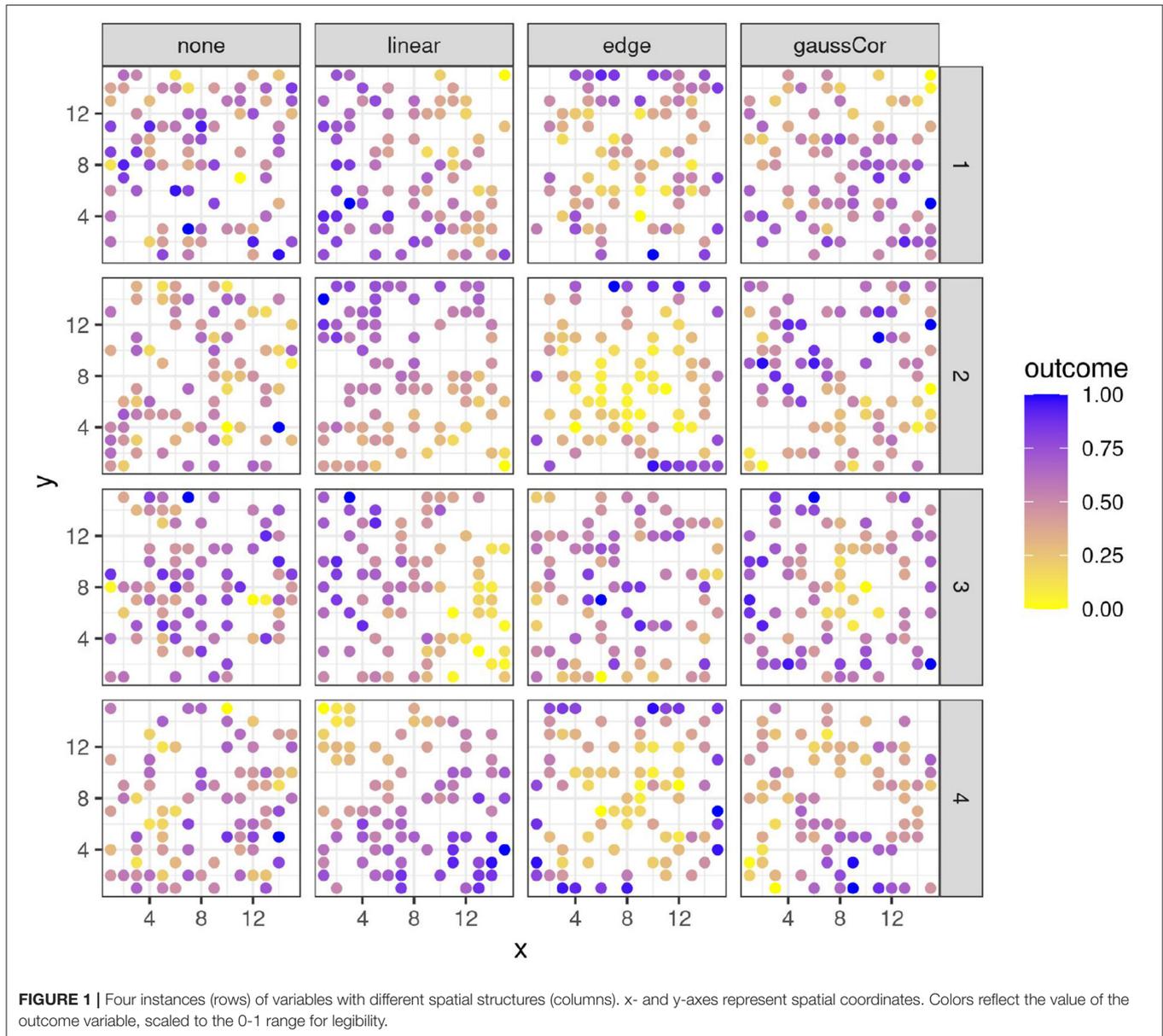
2. MATERIALS AND METHODS

2.1. Simulation Study

2.1.1. Data Generation

Observations are simulated at locations sampled without replacement and with equal probability from an equispaced 15 x 15 grid with interpoint distance 1, indexed by an integer coordinate pair (x, y). Outcome variable A and regressors B_j with $j = 1,..,p$ are generated for samples i=1,...,n at locations (x_i, y_i) as

$$\mathbf{B}_j \sim MVN\left(f_{bj}(\mathbf{x}, \mathbf{y}), \mathbf{\Sigma}_{bj}(\mathbf{x}, \mathbf{y})\right)$$
$$\mathbf{A} \sim MVN\left(f_a(\mathbf{x}, \mathbf{y}) + \mathbf{B}\boldsymbol{\beta}, \mathbf{\Sigma}_a(\mathbf{x}, \mathbf{y})\right) \tag{1}$$

MVN indicates the multivariate normal distribution, f_a and f_b are continuous smooth functions that describe the evolution of the expected values with the location. The $\mathbf{\Sigma}$ matrices describe the covariance structure between the observations as a function of space, whereby the strength of the correlation decreases with the distance between the observations. Hence, specifying a spatial smooth function or a spatial covariance structure are two different ways of engendering spatial structure in the data. All regressors $\mathbf{B}_j$ are drawn independently. Data are drawn for the following scenarios:

1. "None": no spatial structure.

   - $f(x, y) = 0$
   - $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_n$ with I_n the identity matrix of size n and $\sigma^2 = 1$ the residual variance.

2. "Linear": a linear change in baseline.

   - $f(x, y) = (x, y)^t \boldsymbol{\gamma}$ with $\boldsymbol{\gamma}$ a two-dimensional gradient with components drawn uniformly on $[-0.5, 0.5]$ and independently.
   - $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_n$

3. "Edge": observations along the edge have a different mean.

   - $f(x, y) = \frac{\beta_{edge}}{d_{edge}[(x,y)]}$, with $d_{edge}[(x, y)]$ the distance to the closest edge of the field (assumed to be one unit distance away from the border rows) and β_{edge} drawn for each realization from a normal distribution with mean 0 and variance 50.
   - $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_n$

4. "gaussCor": correlation between observations decays as a bell curve.

   - $f(x, y) = 0$
   - $\mathbf{\Sigma} = \sigma^2 \left(\tau \mathbf{I}_n + (1 - \tau)\mathbf{\Sigma}^{gauss}\right)$ with $\Sigma_{ik}^{gauss} = \exp(-(d[(x_i, y_i), (x_k, y_k)]/r)^2)$ and d the Euclidean

FIGURE 1 | Four instances (rows) of variables with different spatial structures (columns). x- and y-axes represent spatial coordinates. Colors reflect the value of the outcome variable, scaled to the 0-1 range for legibility.

distance between two points. $r = 7.5$ is called the range parameter and $\tau = 0.25$ is the nugget parameter.

The parameter values were chosen to qualitatively emulate the spatial patterns observed in real data (Austin and Blackwell, 1980; Langton, 1990; Romani et al., 1993; Haase, 1995; Barmeier and Schmidhalter, 2016; Cruz et al., 2020; Zhou et al., 2021). Some examples of variables generated in these ways are shown in **Figure 1**. All combinations of spatial structures of outcome and regressors were made, whereby all regressors with spatial structure follow the same type of spatial structure per scenario but are otherwise independent. Note that the edge and linear spatial effects increase the variance of a variable, whereas gaussCor decreases its variance (refer to **Supplementary Figure S1**). The number of features was varied

between $p = 1$ (univariate scenario), $p = 50$ (low-dimensional scenario), and $p = 100, 200,$ or 300 (high-dimensional scenario). In the univariate scenario, the design is either observational, with a continuous regressor following a spatial pattern, or experimental, whereby a binary regressor (e.g., treatment vs. control) has a checkerboard pattern across the field. In this latter scenario, an equispaced 18x18 grid was used to allow for partitioning into square subplots with sizes 9, 6, and 3 (refer to **Supplementary Figures S2–S4**), and the 4 different spatial structures are introduced in outcome A only (in the same way as for the 15 x 15 grid). For the training datasets, the sample size was $n = 100$ for the univariate and low-dimensional scenarios and 50 for the high-dimensional scenario. Test datasets were also generated (see below), with the sample size equal to 10 in

all cases (the same size as the left-out folds, refer to Section 2.1.3 below). In the multivariate scenarios, 50% of the regressors were generated with spatial structure, the others without. In both groups of regressors, 20% of the components of $\boldsymbol{\beta}$ were drawn from a zero-mean normal distribution with standard deviations varying between 0.5, 1, and 2 in the low-dimensional scenario and between 0.25 and 0.5 for the high-dimensional scenario, the other components were zero. This setting leads to comparable true R^2 values between low- and high-dimensional scenarios (refer to **Supplementary Figure S5**). For the univariate scenarios, a null setting with $\beta = 0$ was included, as well as an alternative setting whereby β was drawn from a normal distribution with mean 0 and variance 0.25. In the univariate and low-dimensional scenarios, 1,000 Monte Carlo instances were generated, in the high-dimensional scenarios 200. For the INLA models, the number of Monte Carlo instances was limited to 100 in the low- and high-dimensional scenarios for computational reasons.

In a small side simulation of the low-dimensional case, we investigate the role of the number of features p. While fixing the standard deviation of the components of $\boldsymbol{\beta}$ at 1, we let the number of features vary between 10, 20, and 50, with a sample size of $n = 100$.

Even when $\beta_j = 0$, some combinations of the spatial configuration of regressor and outcome lead to non-zero true correlations between their means, for instance, in the case when both experience an edge effect in the observational scenario (refer to **Supplementary Figure S6**), or between edge effect in the outcome and the checkerboard design in the experimental scenario (refer to **Supplementary Figure S7**). This phenomenon is known as spatial confounding. Strictly statistically speaking, this is then not a null scenario, and excess small p-values cannot be seen as a shortcoming of the regression method. Yet, when confined to the two-dimensional space of a field, two variables that are independent can easily exhibit spatial patterns that render them correlated, e.g., two linear gradients in a field are most likely not orthogonal. Therefore, ideally, we would prefer a method that is robust to these kinds of confounded spatial patterns, and we will consider every discovery that results only from this spatial confounding as a false one.

2.1.2. Regression Models
All analyses were performed using the R programming language, version 4.0.3 (R Core Team, 2020). Details on the package versions can be found in **Supplementary Section 5**.

2.1.2.1. Low-Dimensional Scenario
Ordinary least squares was fitted using the *lm* function in the *stats* package. GLS with Gaussian covariance decay assumed was fitted using the *gls* and *corGaus* functions in the *nlme* R-package (Pinheiro et al., 2021). This function iterates between estimating the spatial covariance structure of the residuals ($\hat{\boldsymbol{\Sigma}}$) and fitting an OLS model on **A** and **B** matrices premultiplied by the inverse square root factor of this covariance structure to estimate $\boldsymbol{\beta}$, according to the following model:

$$\hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{A} = \hat{\boldsymbol{\Sigma}}^{-1/2}\mathbf{B}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad (2)$$

with $\epsilon \sim N(0, \sigma^2)$ i.i.d. GAM with 2D thin plate spline with 5 degrees of freedom were fitted using the *gam* function in the *mgcv* package (Wood, 2003). Spatial filtering with the forward selection of spatial eigenvectors or with random effects on the spatial eigenvectors is done with the *esf* and *resf* functions from the *spmoran* package, respectively (Murakami, 2021). Simultaneous autoregressive models were fitted using the *errorsarlm* function in the *spatialreg* package (Bivand and Piras, 2015), using the following two definitions of adjacency. Neighborhoods were defined as either the 8 nearest neighbors ("nearest neighbors"), or all other observations lying within one-fifth of the maximum distance between any two observations ("distance"). In the univariate scenario, we used both versions, for the other analyses, we used only the nearest neighbor version as this one performed best in the univariate models. INLA was applied using the AR(1) model as well as the Matérn model as described by Selle et al. (2019). We used INLA's posterior mean of the β parameter as a point estimate for making predictions and for evaluating parameter estimation and considered the null hypothesis $\beta = 0$ to be rejected when 0 was outside the 95% credible interval. In the univariate scenario, we used both INLA versions, for the other analyses, we used only the Matérn version as this one performed best in the univariate models. When using these methods for making predictions on new datasets, spatial components are omitted from the models, e.g., the thin plate spline for GAM, the spatial eigenvectors for spatial filtering, the neighboring observations for a simultaneous autoregressive model, or the spatial error term for the GLS and INLA Matérn model. These components are only included to obtain a better estimate of $\boldsymbol{\beta}$ but are not included for extrapolation. No feature selection was performed in the low-dimensional scenario; all regressors are included in the models. In the univariate scenario with a checkerboard design, the data are analyzed with and without accounting for row and column effects. Accounting for these effects is done by adding dummy variables for rows and columns (excluding a reference row and column) to the regressor matrix **B**. Note also that in the low-dimensional scenario, no parameter tuning is required, so the parameter estimates are not affected by the CV (refer to Section 2.1.3).

2.1.2.2. High-Dimensional Scenario
For the high-dimensional datasets, a number of variants of the elastic net (EN) implemented by the *glmnet* function in the *glmnet* package are used (Friedman et al., 2010). The mixing parameter α is fixed at 0.5. EN performs feature selection by shrinking some of the coefficients in $\boldsymbol{\beta}$ to zero. We fit the following models:

- The regular EN, without spatial correction.
- A GLS version of EN, which iterates until convergence between

  1. Fitting the EN on the data matrix pre-multiplied by the square root inverse covariance matrix $\hat{\boldsymbol{\Sigma}}^{-1/2}$ as in (2), which estimates $\boldsymbol{\beta}$.
  2. Estimating spatial structure of the residuals $\boldsymbol{\Sigma}$ of this model through restricted maximum likelihood using the

gls function in the *nlme* package, assuming Gaussian covariance decay.

As starting value, the identity matrix is used for $\mathbf{\Sigma}$. This iterative procedure minimizes the following criterion (Friedman et al., 2010)

$$(\mathbf{A} - \beta_0 - \mathbf{B}\hat{\boldsymbol{\beta}})^t \hat{\mathbf{\Sigma}}^{-1} (\mathbf{A} - \beta_0 - \mathbf{B}\hat{\boldsymbol{\beta}}) + 0.5\lambda ||\hat{\boldsymbol{\beta}}||_1 + 0.25\lambda ||\hat{\boldsymbol{\beta}}||_2^2,$$
(3)

where λ is a tuning parameter, β_0 is the intercept, and $\boldsymbol{\beta}$ is the parameter vector of interest, $\mathbf{\Sigma}$ is a nuisance parameter, and $||\hat{\boldsymbol{\beta}}||_1 = \sum_{j=1}^{p} |\beta_j|$ and $||\hat{\boldsymbol{\beta}}||_2^2 = \sum_{j=1}^{p} \beta_j^2$ the L_1 and squared L_2 norms of the $\boldsymbol{\beta}$ parameter vector. Convergence is assumed when the mean squared change in the predictions between two iterations drops below 0.00025.

- An EN model with spatial eigenvectors as regressors in addition to the regular $\mathbf{B}$ regressors. The spatial eigenvectors are found using the *meigen* function in the *spmoran* package (Murakami, 2021).
- A GAM-LASSO model is available in the *plsmselect* package (Ghosal and Kormaksson, 2019). We adapted it to allow the value of the mixing parameter α to be set by the user, which allows for EN to be fitted.

In addition, an INLA Matérn model as in the low-dimensional scenario is fitted. Unlike the previous models, this model is not based on elastic net and does not perform feature selection.

Cross-validation (either blocked or random, see below) is used to tune the penalty parameter λ in the elastic net models, choosing the λ one standard error above the λ with the highest CV R^2, as is the default in the *glmnet* package. Note that since λ is tuned based on different CV paradigms, the test performance on independent test sets may now depend on the paradigm, unlike in the low-dimensional scenario.

2.1.3. Diagnostics

In the univariate scenario, the type I error rate was calculated for the null scenario, and the power to call the association between A and B significant and the mean squared error (MSE) of $\hat{\beta}$ for the alternative scenario. The significance level was set at 5%. For the low and high-dimensional scenarios, methods are evaluated for predictive performance by R^2, which measures the proportion of variance in A explained by the regressors $\mathbf{B}$:

$$R^2 = \frac{Var(A) - Var(A|\mathbf{B})}{Var(A)}.$$
(4)

Generally, the regressors $\mathbf{B}$ are expected to explain only part of the variance in A, yielding R^2 values between 0 and 1. The R^2 is estimated as $\hat{R}^2 = 1 - \frac{\sum_{i=1}^{n}(a_i - \hat{a}_i)^2}{\sum_{i=1}^{n}(a_i - \bar{a})^2}$, with $\hat{a}_i$ the predicted outcome for observation i and $\bar{a}$ the average outcome. When the prediction model is poor, the estimates $\hat{a}_i$ may be such that they increase the variance, leading to negative values of the $\hat{R}^2$ measure. For simplicity, $\hat{R}^2$ is designated by R^2 in the remainder of the document. R^2 was estimated using independent test sets (see below), but also using CV.

It is known that SAC can cause an upward bias in model evaluation by CV and that this effect can be mitigated by

choosing the folds as clusters of nearby observations rather than randomly over the field (refer to **Supplementary Figure S8** for an illustration). This is called blocked or spatial CV and reduces the SAC between the folds (Brenning, 2005; Pohjankukka et al., 2017; Roberts et al., 2017; Meyer et al., 2019; Schratz et al., 2019; Ploton et al., 2020). When CV is used to tune hyperparameters, this blocked CV might also lead to better models by discouraging overfitting, although this could not be demonstrated so far (Schratz et al., 2019). We applied random CV, whereby observations are assigned to folds randomly, and blocked CV, whereby folds are formed by clustering nearby points using the *kmeans* function from the *stats* package. The number of folds was 10 for the low-dimensional and 5 for the high-dimensional scenario, in each case leading to fold sizes of 10. Predictions are mean-centered before evaluating R^2 on the test datasets and in the CV, since we are mainly interested in relative differences within fields rather than in accurate prediction of the baseline per field. Three different R^2 measures were calculated per simulated dataset:

- Cross-validation R^2 (cvR^2), the average of the R^2 values calculated on the left-out folds.
- R^2 on a test dataset of size n=10 (testR^2spatial), generated with the same relation $\boldsymbol{\beta}$ between A and $\mathbf{B}$, and the same spatial structure as the training dataset in regressors and outcome. This means that if, for instance, the spatial structure was linear in the original training dataset, it will also be such in the test dataset, but a new gradient vector $\boldsymbol{\gamma}$ is drawn. Its R^2 then reflects prediction performance on all points of a single, similar plot. A 5 x 5 grid was used to obtain similar spacing between the observations as in the training set.
- R^2 on a test dataset of size $n = 10$ (testR^2none), generated with the same relation $\boldsymbol{\beta}$ between A and $\mathbf{B}$, but without any spatial structure in outcome or regressor. This R^2 reflects the performance of the model over all possible plants from different fields, so without SAC between them.

For evaluation on the test dataset, a model is trained on the training dataset after omitting one of the folds, using the hyperparameter λ optimized in the CV on the entire dataset, such that the size of the training dataset is the same for evaluation on an external test dataset and for CV. Only one test dataset is generated per training set, as the outer Monte Carlo loop of 1,000 instances will average out the test performance over as many test datasets. When the model fitting procedure failed, all R^2 values were set to 0. As a second diagnostic, the MSE of the estimator of $\boldsymbol{\beta}$ is calculated. For the high-dimensional scenario, we also consider the feature selection efficiency. We define *TP* as the number of truly predictive features selected, *FP* as the number of non-predictive features selected and *FN* as the number of truly predictive features not selected, we calculate the sensitivity as $\frac{TP}{TP+FN}$ and true discovery proportion (TDP) as $\frac{TP}{TP+FP}$. We also calculate the proportion of features with spatial patterns among the discoveries, which should be 50% on average if the method does not preferentially select spatial features.

Spatial autocorrelation can be quantified through the Moran's I statistic (Moran, 1950). It can be calculated on the raw outcome

A (marginal Moran's I or I_m) or on the residuals from regressing A on **B** (Moran's I conditional on B or I_c). These measures are defined as:

$$I_m = \frac{n}{s_w} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(a_i - \bar{a})(a_j - \bar{a})}{\sum_{i=1}^{n}(a_i - \bar{a})^2}$$
$$I_c = \frac{n}{s_w} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(a_i - \hat{a}_i)(a_j - \hat{a}_j)}{\sum_{i=1}^{n}(a_i - \hat{a}_i)^2}, \quad (5)$$

with $s_w = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$, $\bar{a}$ denotes the average outcome and $\hat{a}_i$ denotes the outcome predicted by the regression model. The matrix **W** with entries w_{ij} is a weight matrix with zeroes on the diagonal and off-diagonal entries usually decreasing with the distance between observations i and j. Note that the I_c in our case will also capture spatial autocovariance that is included in model components other than the regressors since $\hat{a}_i$ is calculated after omitting all spatial terms, i.e., only $\hat{\beta}$ is used for making predictions. The value of the Moran's I statistic strongly depends on the choice of the weight matrix. Here, we set its off-diagonal elements equal to the reciprocal of the euclidean distance between the points, although this choice is not ideal for detecting edge effects (refer to **Supplementary Figure S9**).

The Moran's I statistic relates the total covariance of each observation with itself $1/n \sum_{i=1}^{n}(a_i - \bar{a})^2$, i.e., the variance (Cov(A, A)=Var(A)), to the short-range weighted spatial autocovariance with nearby observations $\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(a_i - \hat{a}_i)(a_j - \hat{a}_j)}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}}$. Since the weighted spatial autocovariance is always smaller than or equal to the variance, Moran's I statistics are product-moment correlations that lie between –1 and 1 (Day et al., 2004).

The difference between I_m and I_c indicates how the model affects unexplained autocovariance, and could thus be used as a measure of model performance. Yet, because of its definition as a ratio, the Moran's I statistic can be difficult to interpret, as a single, directional change in Moran's I may result from several possible changes in variance and/or short range autocorrelation. As for R^2, I_c was estimated on the test and training residuals, and on the residuals of the left-out folds in the CV. Yet, their comparison is complicated by the fact that the expected value of the Moran's I statistic under the null hypothesis depends on the sample size n: it equals $E(I) = \frac{-1}{n-1}$. It will thus automatically differ when calculated on entire training datasets or smaller left-out folds. As a solution, we subtract its expected value from every Moran's I statistic.

2.2. Case Study

As a case study, we include an investigation on prediction of wheat grain yield and protein content based on Unmanned Areal Vehicle (UAV) spectral reflectance measurements from real farming fields (Zhou et al., 2021). Four fields were measured in two different years (Fields 1 and 2 in 2018 and Fields 3 and 4 in 2019, with 90 and 101, and 68 and 68 samples, respectively) for the same crop (winter wheat) and cultivating conditions. For the first two fields, plant samples were collected along 3 lateral and one longitudinal transects in the field. For the other two fields, plant samples were collected from 10 m × 15 m grid points across

each field (refer to **Supplementary Figure S10**) and measured for yield and protein content (**Supplementary Figure S11**). All fields were also surveyed with a camera mounted on a UAV. As usual, to eliminate confounding factors such as atmospheric effects and sun angle, features in the form of vegetation indices and spectral ratios were extracted from the resulting images prior to model fitting. Because of high multicollinearity between these indices (refer to **Supplementary Figures S12–S16**), for each field, only the set of most important principal components that together explain at least 99% of the variance are included as regressors (3 components for fields 1–3 and 4 components for field 4), refer to **Supplementary Figure S17**. Low-dimensional prediction models as introduced in Section 2.1.2.1 are trained on all fields separately, and predictive performance is estimated using 5-fold CV. The CV was repeated 20 times and resulting cvR^2s were averaged. Next, the trained models were validated by predicting yield and protein content for other fields, mean-centering these predictions, and comparing them to mean centered observed values of other fields. For validating the models on other fields, we used the entire training dataset to fit the model, rather than leaving out 1 fold as in the simulation study.
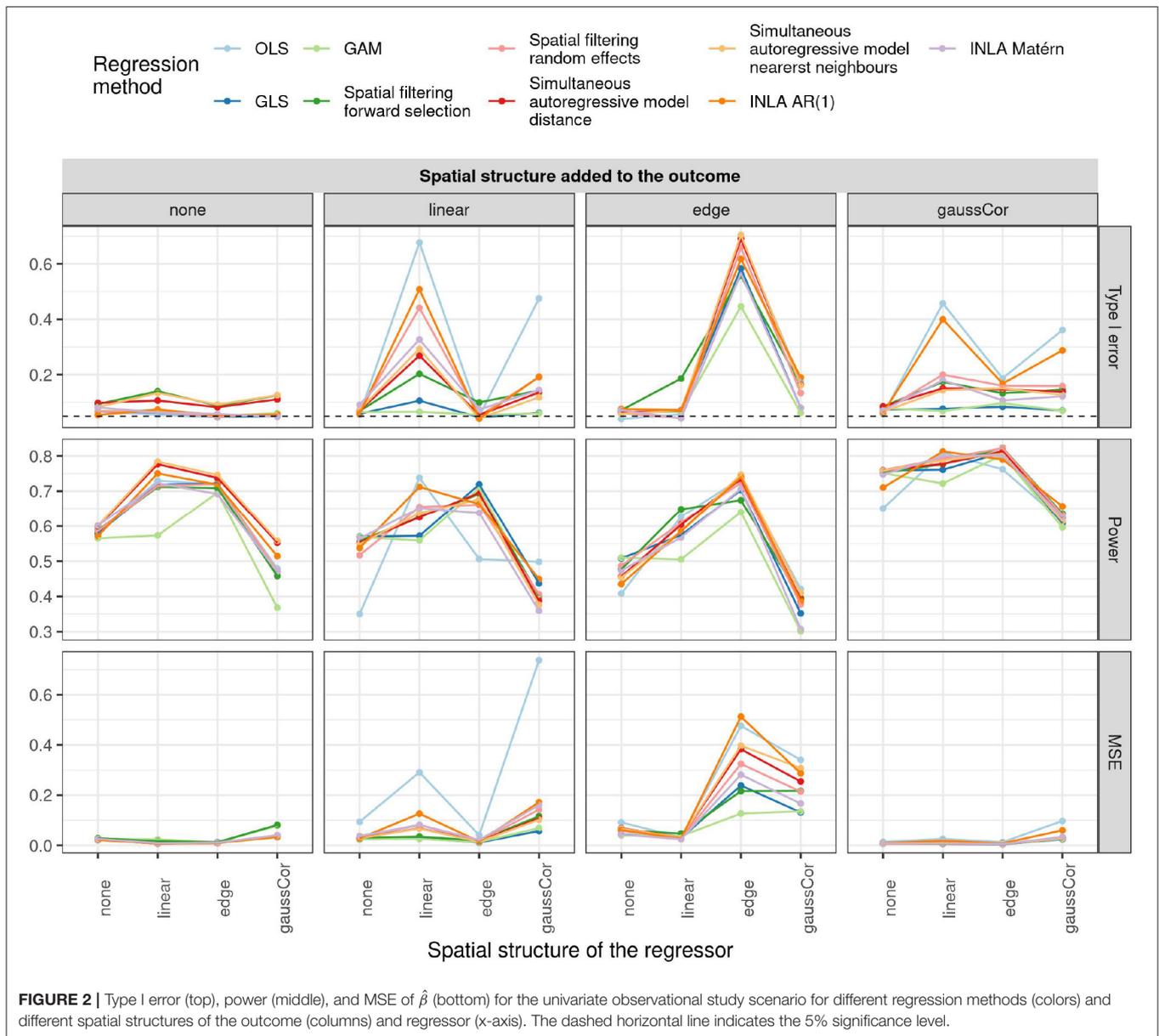
3. RESULTS

The performance of different spatial regression methods was investigated in simulations for two univariate scenarios and for multivariate low- and high-dimensional scenarios. Next, the results were compared to the performance of the methods on a field trial dataset on winter wheat (Zhou et al., 2021).

3.1. Univariate Scenario

The type I error, power, and MSE of $\hat{\beta}$ for the observational study scenario are shown in **Figure 2**. OLS, spatial filtering, INLA Matérn, and the autoregressive models (simultaneous and INLA) generally struggle the most to control the type I error in presence of spatial effects. On the other hand, GLS and GAM are most robust to spatial confounding, as they generally control the type I error rate at around 5–10%. However, when both outcome and regressor experience an edge effect, no method manages to detect the spatial confounding and all methods call the association significant. Also, combinations of linear and/or Gaussian SAC patterns in the regressor and outcome generally produce high type I error rates, except when using GLS or GAM.

For the checkerboard design, the inflation of the type I error is worst for the case of Gaussian correlation in the outcome, but this inflation is in all cases greatly reduced by using small subplots (**Figure 3**). Spatial filtering with forward selection generally has a large type I error, presumably because it executes multiple statistical tests in the selection of eigenvalues but only reports the final p-value, without accounting for the previous tests. Also OLS, GLS, spatial filtering with random effects, and the autoregressive models have inflated type I error in some scenarios. Note that inflation of the type I error due to edge effect is worst for the intermediate subplot size of 6, especially for INLA AR(1), GLS, and OLS without correction for row/column effects. The reason is that for this design (refer to **Supplementary Figure S3**), the

FIGURE 2 | Type I error (top), power (middle), and MSE of $\hat{\beta}$ (bottom) for the univariate observational study scenario for different regression methods (colors) and different spatial structures of the outcome (columns) and regressor (x-axis). The dashed horizontal line indicates the 5% significance level.

edge is spatially confounded with the treatment effect because the number of subplots is uneven in this case and either the treatment or the control plots tend to be located toward the edge (refer to **Supplementary Figure S7**). This result suggests that checkerboard designs are best made with even numbers of subplots to avoid spatial confounding with edge effects. It also explains why there is no inflation of the type I error for the edge effects with other subplot sizes and all linear effects: they are orthogonal to the checkerboard design pattern and hence do not cause spatial confounding. Results on the power and MSE of all methods can be found in **Supplementary Figures S18, S19**.

3.2. Low-Dimensional Scenario

An overview of the results for the low-dimensional scenario is shown in **Figure 4**, more results can be found

in **Supplementary Figures S20–S23**. When the outcome exhibits an edge effect, the testR^{2}spatial is generally lower than the testR^{2}none, probably because the extra variability of the edge effect cannot be explained by the regressors. When the regressors have a linear or Gaussian spatial structure, random CV overestimates R^2 on unseen test datasets, whereas blocked CV is more accurate. OLS and simultaneous autoregressive models have poor performance on test datasets when there is a linear or edge structure in the outcome, but a different spatial structure in the regressors. For linear structure in the outcome, spatial filtering and INLA Matérn exhibit intermediate performance and GAM and GLS perform best. Spatial filtering and INLA Matérn as well as GAM and GLS perform when there is a Gaussian autocorrelation or edge structure in the outcome;
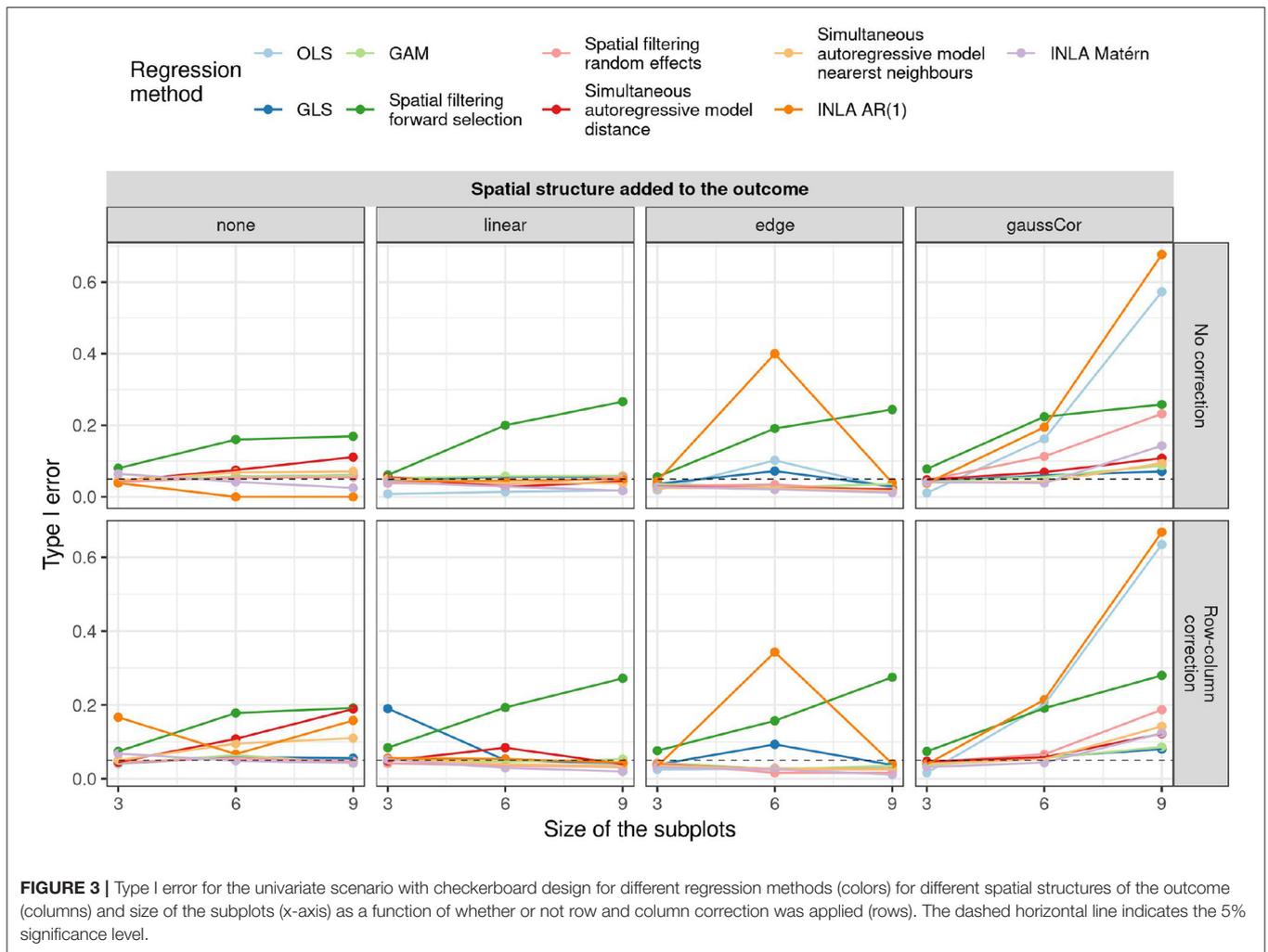
FIGURE 3 | Type I error for the univariate scenario with checkerboard design for different regression methods (colors) for different spatial structures of the outcome (columns) and size of the subplots (x-axis) as a function of whether or not row and column correction was applied (rows). The dashed horizontal line indicates the 5% significance level.
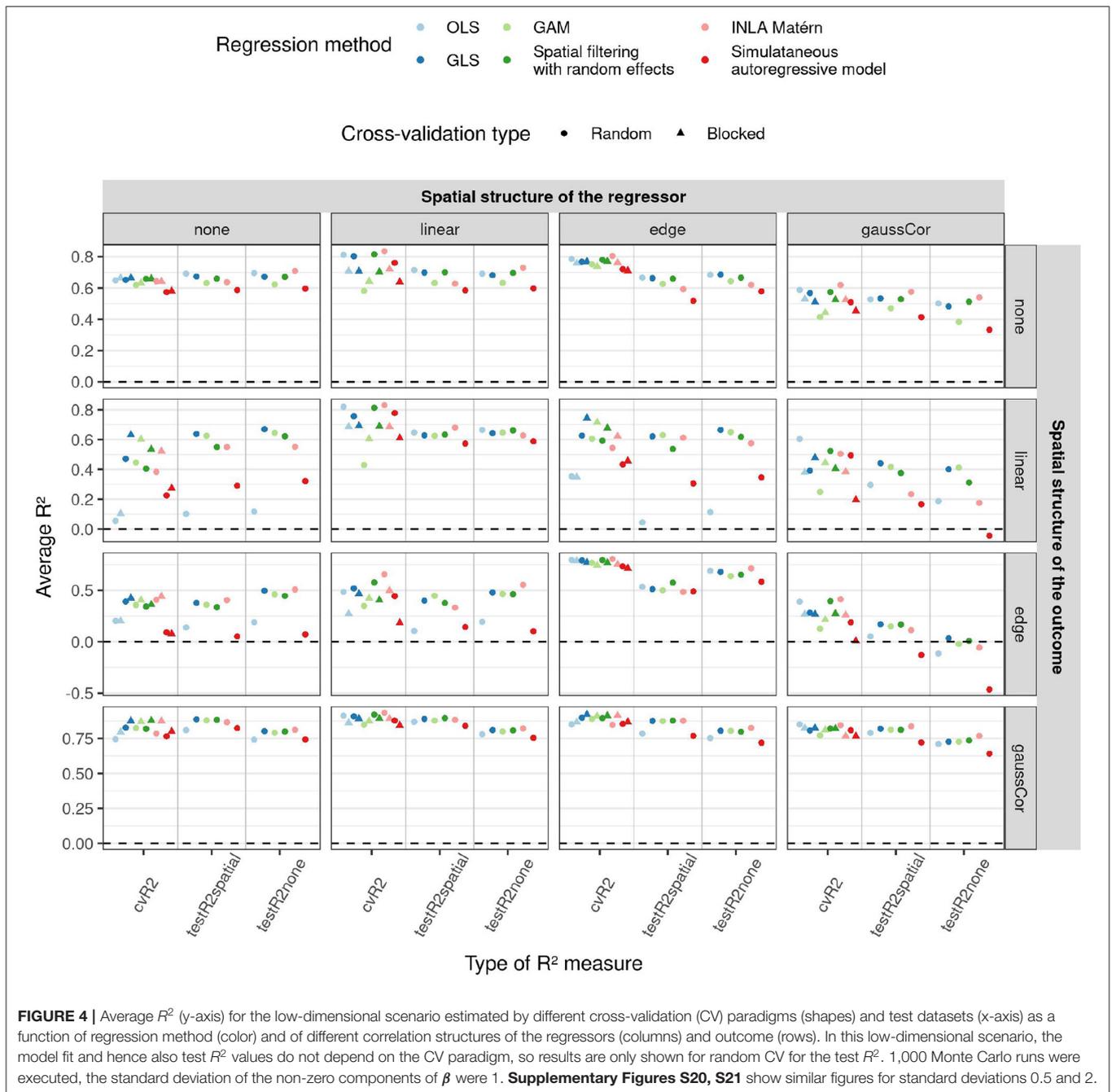
GAM performs poorly in absence of spatial structure in the outcome.

We also ran simulations investigating the role of the number of features. **Supplementary Figures S24–S26** show that the difference between the regression methods is generally smaller when built on fewer features. The performance of the simultaneous autoregressive models deteriorates as the number of features increases, even in absence of any spatial pattern. These graphs also reveal that a higher number of features does not always lead to better predictions, as the estimation of the corresponding parameters becomes harder.

Interestingly, despite their poor performance, OLS and simultaneous autoregressive models have the lowest SAC in the residuals of the training data, as measured by the Moran's I statistic (**Figure 5**). Although the spatial structure cannot be entirely explained by the regressors, i.e., when there is an independent spatial pattern in the outcome, the values of the Moran's I statistic are often close to their expected value under the null hypothesis of no SAC (and hence on the plot close to zero after subtracting the expected value). Yet, when the corresponding models are applied to a new test dataset with

spatial structure, there is little difference between the methods in terms of residual spatial autocorrelation. These results suggest that methods ignoring spatial structure, like OLS, try to capture realizations of the spatial effects using the available covariates, yielding low residual SAC in the training data. This gives the false impression of explaining all spatial structures in the outcome through the spatial patterns in the regressors, but this is a case of overfitting the spatial pattern. Just like Moran's I on the test datasets, blocked CV Moran's I was in many cases similar among the methods, indicating that blocked Moran's I CV is a better predictor for differences between the methods in residual spatial autocorrelation than training or random CV Moran's I. For instance, in presence of linear effects in the regressor, GAMs often have large residual SAC in the training dataset and in random CV, but not in the blocked CV or test data with spatial structure.

In summary, GLS generally works best across different SAC scenarios, even when its assumptions on spatial error structure are not entirely fulfilled. An exploration of how GLS accommodates linear and edge effects is given in **Supplementary Section 3**.
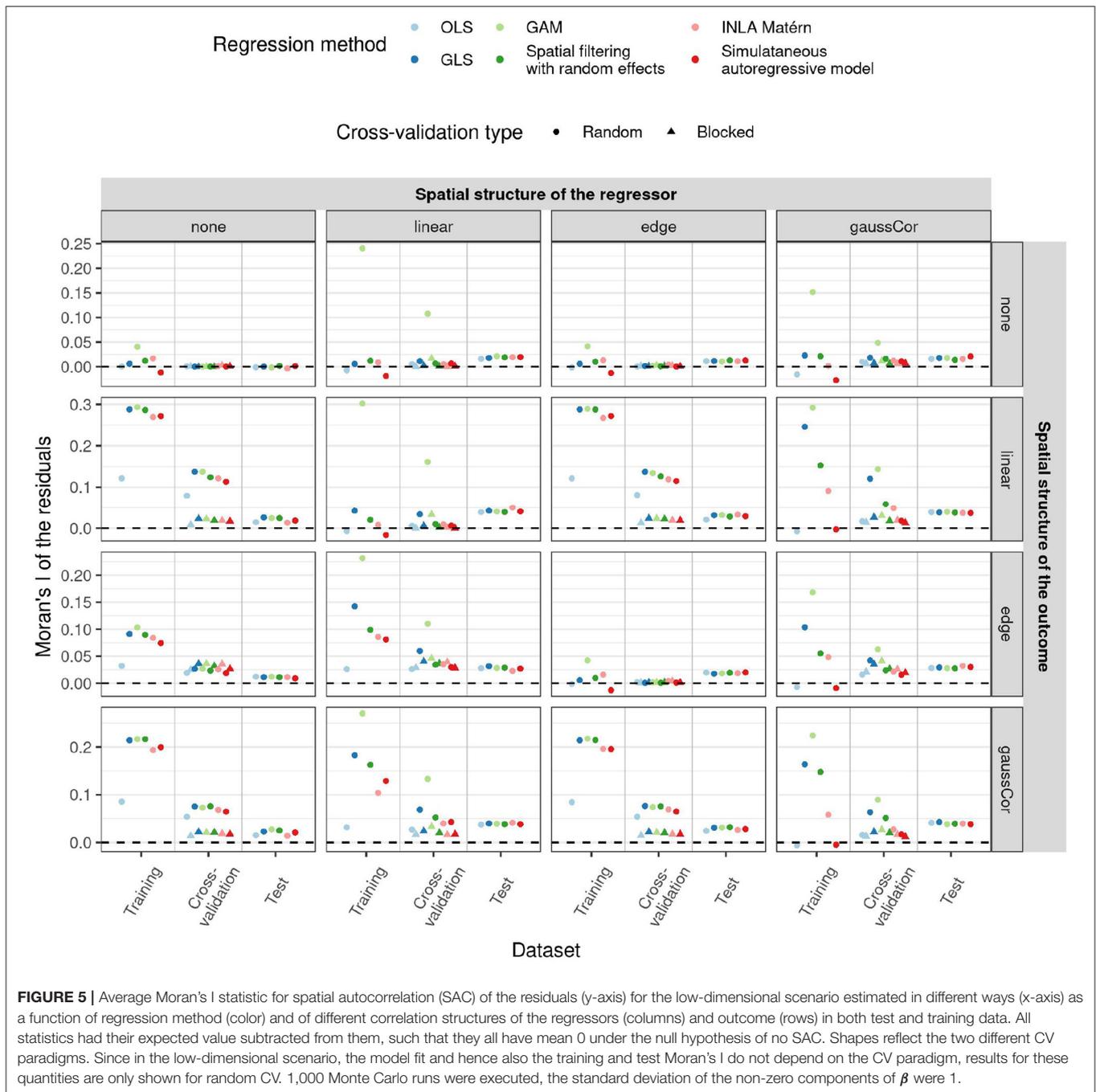
FIGURE 4 | Average R^2 (y-axis) for the low-dimensional scenario estimated by different cross-validation (CV) paradigms (shapes) and test datasets (x-axis) as a function of regression method (color) and of different correlation structures of the regressors (columns) and outcome (rows). In this low-dimensional scenario, the model fit and hence also test R^2 values do not depend on the CV paradigm, so results are only shown for random CV for the test R^2. 1,000 Monte Carlo runs were executed, the standard deviation of the non-zero components of β were 1. **Supplementary Figures S20, S21** show similar figures for standard deviations 0.5 and 2.

3.3. High-Dimensional Scenario

An overview of the results for the high-dimensional scenario is shown in **Figure 6**, more results can be found in **Supplementary Figures S27–S38**. The performance of the prediction models for high-dimensional data is seen to decrease with the number of features (as shown in **Figure 6**), just like in the low-dimensional scenario (**Supplementary Figure S25**). An exception is INLA Matérn, which performs worse than the other methods at low numbers of regressors ($p = 100$), but better when the number of regressors is high ($p = 300$). However, its model

fitting procedure failed for around 20% of the Monte Carlo instances in all settings.

The feature selection differs much more between the different methods than predictive performance: GLS EN has higher sensitivity in selecting correct predictors in all scenarios, whereas GAM EN has lower sensitivity than the other methods (**Supplementary Figure S31**). On the other hand, GLS EN has a slightly lower TDP in some scenarios, whereas GAM EN generally has a higher one (**Supplementary Figure S32**). Overall, the TDP and especially the sensitivity decrease with the number

FIGURE 5 | Average Moran's I statistic for spatial autocorrelation (SAC) of the residuals (y-axis) for the low-dimensional scenario estimated in different ways (x-axis) as a function of regression method (color) and of different correlation structures of the regressors (columns) and outcome (rows) in both test and training data. All statistics had their expected value subtracted from them, such that they all have mean 0 under the null hypothesis of no SAC. Shapes reflect the two different CV paradigms. Since in the low-dimensional scenario, the model fit and hence also the training and test Moran's I do not depend on the CV paradigm, results for these quantities are only shown for random CV. 1,000 Monte Carlo runs were executed, the standard deviation of the non-zero components of β were 1.

of features, indicating that the task of feature selection becomes more difficult as the dimensionality of the dataset grows. This also explains the decrease in model performance with the number of features, despite the fact that there are more truly predictive features and thus theoretically a higher R^2 (refer to **Supplementary Figure S5**).

Interestingly, in the presence of spatial patterns in the regressors, regular EN selects more than the expected 50% of features with a spatial structure (refer to **Supplementary Figure S33**), the so-called "red-shift" (Lennon, 2000). EN with spatial filtering only suffers from this red-shift

when there are linear gradients or edge effects in the data. GLS EN appears less sensitive to red-shift, selecting around 50% of spatial features in case of linear patterns or Gaussian correlation in the regressors. Yet, in the case of edge effects in the regressors, GLS EN also selects excess spatial features. GAM EN on the other hand, often selects less than 50% spatial features.

Generalized additive model EN generally has a higher SAC in the residuals than competing methods when modeling training data or left-out CV folds, in particular when there are linear trends in the regressors. INLA Matérn on the other hand has
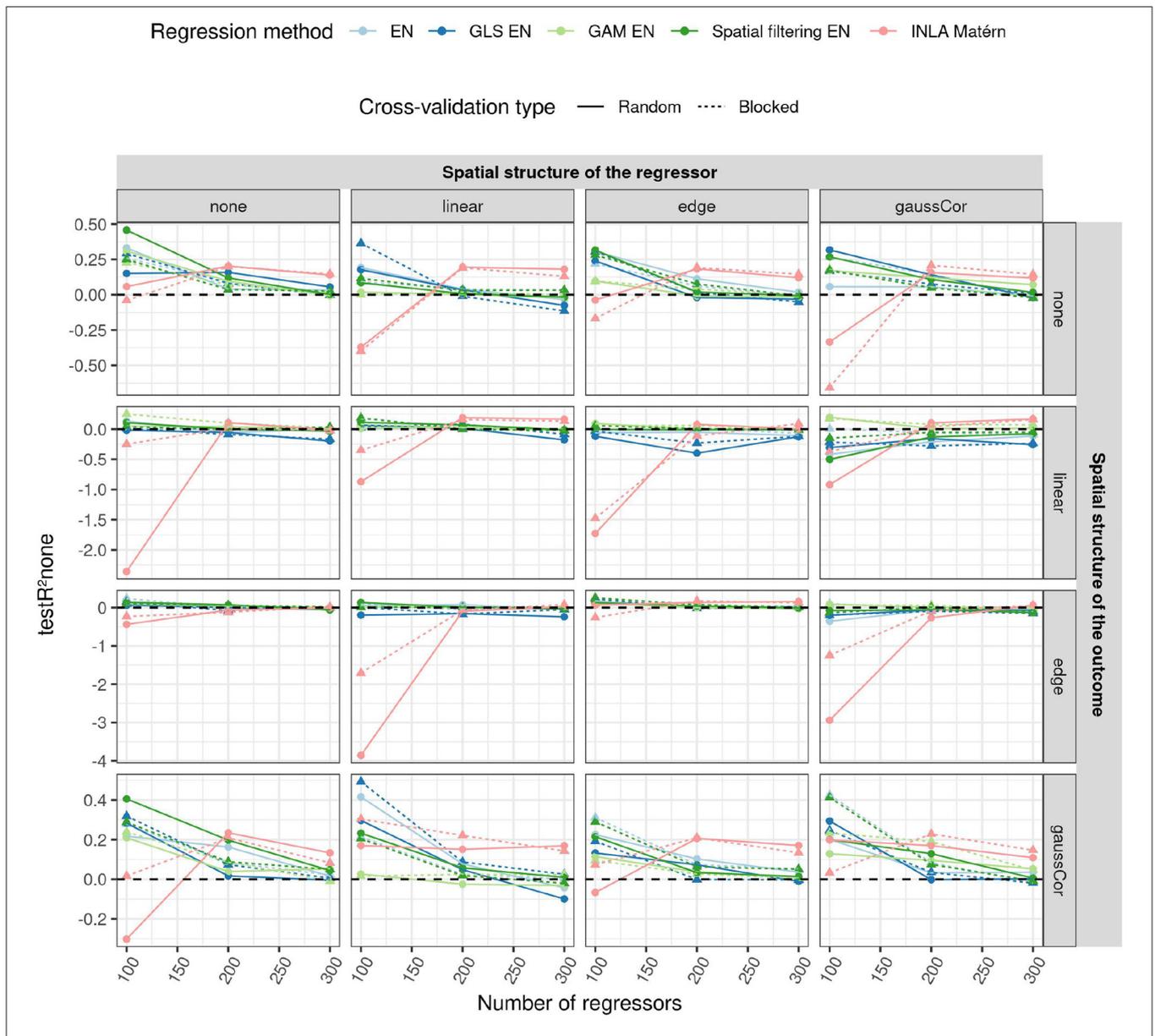
FIGURE 6 | Average R^2 on an independent test dataset without spatial structure (testR^{2}none, y-axis) for the high-dimensional scenario when using different CV paradigms (shapes and linetypes) as a function of the number of regressors (x-axis), EN version (colors), the correlation structure of the regressors (columns) and of the outcome (rows). In the high-dimensional case, CV is used to tune the penalty parameter, so the model fit may depend on the CV paradigm. For this reason, test R^2 is shown for both CV paradigms in this case. The standard deviation of the non-zero components of the β parameter was 0.5. The horizontal dashed line indicates an R^2 of 0.
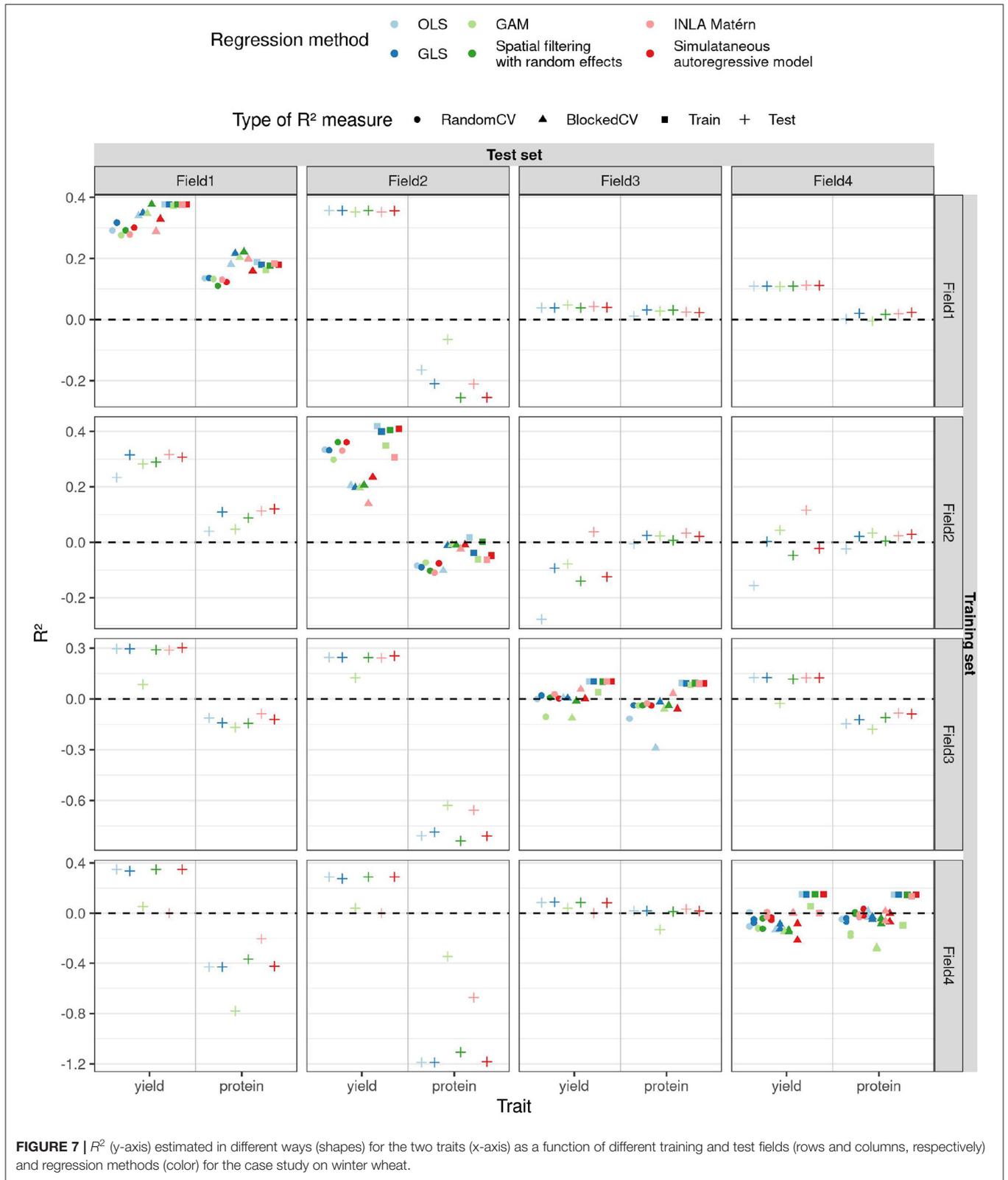
a lower residual SAC when modeling training data or test data with linear trends in the outcome or regressors (refer to **Supplementary Figures S34–S37**).

3.4. Case Study: Predicting Winter Wheat Yield Based on Multispectral Measurements

3.4.1. Data Description

The scatterplots and Pearson correlations in **Supplementary Figures S12–S15** indicate strong

multicollinearity between the multispectral variables, but less correlation between yield and especially protein content on the one hand and the multispectral variables on the other. PCA biplots in **Supplementary Figure S16** suggest a similar covariance pattern between all variables over the 4 different fields, although *nir* and *evi*2 are less correlated with the other variables in fields 3 and 4 than in fields 1 and 2. Average yields differ between the fields, as shown in **Supplementary Figure S11**, hence a correction for a different baseline is needed. Field 2 has the strongest autocorrelation in almost all variables, followed by field 1 (refer to **Supplementary Figure S39**).

FIGURE 7 | R^2 (y-axis) estimated in different ways (shapes) for the two traits (x-axis) as a function of different training and test fields (rows and columns, respectively) and regression methods (color) for the case study on winter wheat.

3.4.2. Prediction Model Performance

Predicted vs. observed yield and protein content are shown in **Supplementary Figures S40, S41**. The CV, training, and test R^2 are shown in **Figure 7**, estimated β parameters of the principal components with standard errors can be found in **Supplementary Figure S42**. We see that the standard errors of the parameter estimates increase with the order of the principal component, but do not really depend on the method. An exception is the INLA Matérn model trained on field 4, which has small estimates and large standard errors. The sums of the products of the parameters with principal component loadings $\beta \mathbf{u1}$, with $\mathbf{u}$ the matrix of principal components included in the model and $\mathbf{1}$ a column vector of ones, are plotted in **Supplementary Figure S43**. Overall, we find lower values of R^2 than reported by the authors (Zhou et al., 2021). This is likely because we 1) split the data into separate fields, rather than scrambling data from all fields together, 2) use smaller training datasets and 3) do not ensure equal variances in training and test set. Yet, our objective was to estimate generalisability of the models between fields, which is different from the objective of the original paper.

For yield, we observe that when training on field 2, the field with the strongest SAC, the blocked CV R^2 estimate is lower than the random one (as shown in **Figure 7**). When trained on this field, OLS performs poorly on test sets, as its predictions are often too extreme (refer to **Supplementary Figure S40**), and also its parameter estimates are more extreme than for the other methods (refer to **Supplementary Figure S42**). Also for field 2, OLS has a smaller SAC of the residuals of the training data than the other methods (refer to **Supplementary Figure S44**). These results are in accordance with the results of our simulation study: SAC deteriorates the performance of OLS and causes blocked and random CV R^2 estimates to diverge, but residuals of the training data of OLS are not spatially autocorrelated. The GAM models for yield trained on fields 3 and 4 perform worse on other fields than the other methods, their predictions being too extreme for the model trained on field 3 and too average for the model trained on field 4 (refer to **Supplementary Figure S40**). The corresponding parameter estimates of GAM follow the same pattern (refer to **Supplementary Figure S42**). This had been predicted by CV (for field 4 only by random CV) and is reminiscent of the poor performance of GAM on data with Gaussian correlation found in the simulation study. The INLA Matérn model trained on field 2 generalizes best to fields 3 and 4, but on the other hand, the INLA Matérn model trained on field 4 has low training and test R^2.

For models predicting protein content, the model performances are often worse, in concordance with the authors who noticed this phenotype was harder to model using linear models (Zhou et al., 2021). Prediction is also complicated by large outlying values in fields 3 and 4 (refer to **Supplementary Figure S41**). Yet, also here, OLS performs slightly worse than other models on test fields when trained on field 2, as predicted by blocked CV R^2 only. Blocked CV on field 3 also predicted a lower R^2 for OLS, but this is not seen in the test R^2. For fields 1 and 2, the blocked CV R^2 is higher than random CV R^2 for all model types, but this is only accurate

according to the test R^2 values for the protein content models trained on field 2. The simultaneous autoregressive model, INLA Matérn, and spatial filtering methods perform overall similarly to GLS. Yet our simulation results suggest that the simultaneous autoregressive model performs well in this case thanks to the low number of features, but that its performance would deteriorate as the number of features increases (refer to **Supplementary Figure S24**). The good performance of the spatial filtering method suggests that the setting of this case study is more akin to the Gaussian correlation or edge effect scenario than to linear effects, as could be expected from **Supplementary Figure S10**.

Looking at generalisability of the yield models, the yield is best predicted for fields 1 and 2, regardless of which field the models were trained on. On the other hand, the yield on fields 3 and 4 is harder to predict. The training R^2 on fields 3 and 4 is even lower than the test R^2 when the corresponding models are applied to fields 1 and 2, which indicates that simply more variability in yield in fields 1 and 2 can be explained by the multispectral variables measured. This is a different situation from the simulation study, where the residual variance σ^2 was kept constant, allowing us to use CV R^2 as an estimator for test R^2. Nevertheless, CV was able to predict relative differences in performances in the case study and can be a great help in this respect. Despite higher noise levels in fields 3 and 4, the models trained on them are not less accurate than the models trained on field 1. The least accurate models are the ones trained on field 2, illustrating how SAC deteriorates model estimation, even when suitable spatial regression models are used.

4. DISCUSSION

In this study, we compare a number of regression methods for data with spatially autocorrelated variables. Simultaneous autoregressive models were only found to work well when the number of regressors is low. Spatial filtering methods performed well in the case of unstructured SAC, but struggled to deal with structured forms of SAC such as linear gradients and edge effects. Generalized additive models (GAMs) account for SAC by explicitly modeling the dependence of the outcome on the location, which yields good insights into the spatial structure of the outcome variable. On the downside, GAMs have a high efficiency cost when there is no long-range spatial dependence in the data. Moreover, GAMs tend to absorb too much of the spatial structure of the outcome in their smooth spatial term, thereby underrating the ability of the available regressors to explain this spatial structure. This overcorrection can lead to a veritable "blue-shift," whereby regressors with spatial signals are less likely to be recognized as associated with the outcome. The INLA model based on a Gaussian random field with Matérn covariance structure performed reasonably well in both low- and high-dimensional scenarios. When the dimensionality of the data is much higher than the sample size, the INLA Matérn model even outperforms the elastic net models in terms of prediction accuracy. On the downside, its Bayesian model-fitting algorithm may fail to converge, and no

TABLE 1 | Best methods for different scenarios (first row) and criteria (second and third rows).

Univariate	Low-dimensional	High-dimensional			
Hypothesis testing	Prediction	Prediction	Feature selection		
			Sensitivity	TDP	Robustness to red-shift
GLS, GAM	GLS	EN for moderately high dimensions, INLA for very high dimensions	GLS EN	GAM EN	GLS EN

TDP, true discovery proportion; GLS, generalized least squares; GAM, generalized additive model; EN, elastic net. See text for details.

feature selection in the high-dimensional scenario is provided. The GLS method was found most robust in accounting for different forms of SAC, in both low and high-dimensional settings. In the low-dimensional scenario, GLS comes with a slight efficiency loss compared to OLS when there is no spatial structure, but outperforms OLS by a large margin in most settings with spatial correlation. For the high-dimensional setting, we newly combined GLS estimation of the covariance structure of the residuals with a penalized least squares method (elastic net). Although we found no improvement in terms of predictive performance, and in some scenarios even a slight decrease, we have shown how it improves upon regular elastic net with respect to feature selection. In the presence of spatial effects, GLS elastic net has higher sensitivity in detecting good predictors and is less sensitive to red-shift (selecting excess features with spatial patterns) than competing methods. This confirms the trade-off between predictive performance and feature selection as described by Wang and Zhu (2009): there are no single best methods for all scientific aims. As a recommendation to the reader, the methods that performed best in the various simulation scenarios are summarized in **Table 1**.

In summary, for most purposes, we recommend the use of GLS to account for SAC effects in regression models. GLS has the flexibility to accommodate many forms of autocorrelation in its error structure (refer to **Supplementary Figure S45**), e.g., anisotropic effects whereby the SAC does not decay equally fast with distance in all directions, or temporal autocorrelation. GLS can be combined with many statistical and machine learning methods through a simple iterative loop, even when this increases computation time (refer to **Supplementary Section 4**). Note that the iterative procedure we propose is different from simple preconditioning on the marginal spatial covariance structure of the outcome, which has been condemned before for rendering the design matrix ill-conditioned (Jia and Rohe, 2015). In our proposal, as in most regression models, the distribution of the outcome (including its covariance structure) conditional on the regressors is of interest. GLS could also be extended to generalized linear models (GLMs) through their formulation as iteratively reweighted least squares. Yet despite GLS's good performance, the estimation problem of its spatial covariance matrix is so ill-posed that its components (range, nugget, and residual variance) are strongly correlated and hence very variable (refer to **Supplementary Section 3**).

Since GLS and some other spatial methods account for spatial structure in the errors, the residuals of these models

may remain spatially autocorrelated, which is not problematic. The main purpose of a regression method is to estimate a relationship between regressors and outcome, which helps to explain some of the total variance of the outcome. The degree to which this succeeds can be measured by R^2. Yet, a low R^2 does not necessarily imply that the model is wrong, or that its parameter estimates are inaccurate, as a low R^2 may also result from missing important covariates or strong background noise in the training data. A predictive model trained on a noisy dataset with low training R^2 may perform much better on a test dataset with low noise levels, as was demonstrated in the case study on winter wheat. Analogously, as a side effect of model fitting, regression methods may explain part of the marginal spatial autocovariance (as for instance measured by the Moran's I statistic Moran, 1950) in the outcome variable through the spatial patterns found in the regressors, leading to a reduced spatial autocovariance in the residuals. Yet, this is not the primary purpose of the regression method, and a large residual spatial autocovariance does not imply inaccurate estimation of the relationship between outcome and regressors. The residual spatial autocovariance may, just like residual variance, result from missing covariates or spatially autocorrelated noise. If almost all spatial autocovariance can be explained by the regressors (corresponding to the simulation scenarios without additional spatial patterns applied to the outcome), methods ignoring spatial structure work fine and are slightly more efficient than spatially aware methods, at least in the low-dimensional scenario. However, when not all spatial autocovariance can be explained by the regressors (the simulation scenarios with additional spatial patterns applied to the outcome), a method like OLS that ignores spatial structure will nevertheless attempt to fulfill its assumption of i.i.d. errors. This can lead to small residual spatial autocovariance in the training dataset, but also to variable or biased parameter estimates and preferential selection of regressors with spatial patterns (the "red-shift" Lennon, 2000). Hence, the absence of spatial autocovariance in the training residuals is not a guarantee that the spatial structure of the data has been well-modeled (Beale et al., 2010; Roberts et al., 2017), and the model may still fail to explain spatial autocovariance in a new dataset (refer to **Figure 5**). Since it is hard to predict whether all spatial autocovariance is explicable by the regressors, and the efficiency cost of accounting for spatial structure is low, we recommend always using spatially aware methods when analyzing datasets with potential for spatial structure, such as field trials. Finally, note that the opposite occurrence is indeed problematic: when spatially naive methods like OLS

nevertheless have residual SAC, this points at problems with the model fit.

Past comparative simulation studies generated data in an unstructured way (Wang and Zhu, 2009; Beale et al., 2010; Alesso et al., 2019; Rocha et al., 2019, 2021; Mao et al., 2020; Harisena et al., 2021), ignoring the fact that spatial patterns in real fields often take on discernible shapes such as edge effects or linear gradients (Austin and Blackwell, 1980; Langton, 1990; Romani et al., 1993; Haase, 1995; Sarker and Singh, 2015; Barmeier and Schmidhalter, 2016; Cruz et al., 2020; Zhou et al., 2021), and despite linear trends and row- and column effects often being included in the analysis model of real data as well (Singh et al., 2003; Sarker and Singh, 2015). In this study, we have included the linear trend and edge effects scenarios as well and found that they can have a profound effect on model performance. Edge effects are difficult to account for, inflating type I error in hypothesis tests and deteriorating model performance on test fields. Hence, they should preferably be avoided by adequate field trial design. Linear gradients in outcome and regressors are also challenging, as they can easily become spatially confounded in the two-dimensional space of a field. Yet, these linear effects can be captured rather well by spatial regression methods such as GAM and GLS. In theory, one could also account for linear or edge effects in the mean model of an OLS regression when their presence is known beforehand, or obvious from the data. Yet in practice, these effects may not be easily detectable, and combinations of different kinds of spatial patterns may be present in the data, such that robust spatial regression methods seem to be a safer choice.

In the setup of our simulation study, we let linear and edge effects increase the variance of a variable. Whether this assumption is correct is difficult to verify. A higher variance in the outcome variable complicates its prediction, whereas a higher variance in a regressor variable renders the model estimation more precise. More commonly used autocovariance simulation schemes make nearby observations more similar, which reduces the information content of the dataset and increases the variance of parameter estimators. The effect on feature selection is similar though: both short range positive SAC and long range linear gradients and edge effects cause excess selection of spatial features in hypothesis testing and penalized regression.

Correction for spatial effects in field trials is common, but often only occurs between discrete plots of checkerboard designs (Lado et al., 2013; Elias et al., 2018; Alesso et al., 2019; Rocha et al., 2019; Mao et al., 2020; Harisena et al., 2021). Yet in our own simulations, we see that such corrections have only limited effect, and SAC within and across discrete plots can still inflate type I errors and estimator variability. It has been demonstrated that the use of more, smaller subplots can combat this (Alesso et al., 2019), but may involve more work and is only possible at the design stage. On top of using small subplots, we, therefore, recommend spatially aware regression methods such as GLS and GAM for the analysis of checkerboard designs, as they yield good power while controlling type I error rate.

It is known that CV can misestimate model performance on an independent test dataset (Rocha et al., 2018). On the one hand, the model may overfit the training set, and on the other hand, noise levels may differ between training and test set, rendering measures such as mean squared error (MSE) of the predictions and R^2 hard to compare. Yet, this problem may be exacerbated by spatial patterns in the training data. SAC causes information leaks between the training set and left-out fold, violating the independence assumption and causing estimates of model performance to be overoptimistic. It has been shown that this problem can be mitigated, but often not eliminated, by blocked CV, i.e., choosing the folds as blocks of neighboring observations (Brenning, 2005; Pohjankukka et al., 2017; Roberts et al., 2017; Meyer et al., 2019; Schratz et al., 2019; Ploton et al., 2020). We confirm that blocked CV is better able to discriminate between competing methods in the presence of SAC, and is similar to random CV in absence of such autocorrelation. Hence, we recommend blocked CV as a simple and widely applicable solution in all scenarios where SAC may be present. Other corrections for CV on dependent data have been proposed (Rabinowicz and Rosset, 2020), but these are restricted to a smaller class of methods and were not tested here. Furthermore, it has been hypothesized before that a change in CV paradigm could affect hyperparameter tuning (Schratz et al., 2019), but this was not found in our study. Another problem with SAC arises when the test dataset also exhibits spatial patterns that are not completely explicable by the regressors. This causes additional deviations, with a spatial pattern, of the observed values from the expected ones (Rocha et al., 2018).

In summary, we believe that field trials, either observational or experimental, are a great way to learn relationships between variables if the following points are taken into consideration. In simulation studies for spatial data, we recommend the inclusion of more realistic spatial structures such as edge effects and linear gradients. In checkerboard designs, the use of many, small subplots mitigates spatial confounding. For the analysis of data with spatial signatures, we support the use of blocked CV. Moreover, we suggest the inclusion of spatial covariance matrices in regression models, including high-dimensional ones, through GLS, and warn against model evaluation through residual SAC of the training dataset only.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The dataset is available from Zhou et al. upon request. Requests to access these datasets should be directed to takashit@gifu-u.ac.jp.

AUTHOR CONTRIBUTIONS

SH and SD conceived the study. SH performed the analyses. SH wrote the manuscript with input from SD and SM. SM supervised the work. All authors interpreted the results.

FUNDING

of the Research Foundation-Flanders (FWO, grant 1146319N).

ACKNOWLEDGMENTS

We are grateful to Takashi Tanaka for kindly providing data and explanation of the winter wheat dataset. We also thank Véronique Storme for her suggestion to include the INLA method in the comparison.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.858711/full#supplementary-material

REFERENCES

Alesso, C. A., Cipriotti, P. A., Bollero, G. A., and Martin, N. F. (2019). Experimental designs and estimation methods for on-farm research: a simulation study of corn yields at field scale. *Agron. J.* 111, 2724–2735. doi: 10.2134/agronj2019.03.0142

Austin, R. B., and Blackwell, R. D. (1980). Edge and neighbour effects in cereal yield trials. *J. Agric. Sci.* 94, 731–734. doi: 10.1017/S0021859600028720

Barmeier, G., and Schmidhalter, U. (2016). High-throughput phenotyping of wheat and barley plants grown in single or few rows in small plots using active and passive spectral proximal sensing. *Sens. Basel* 16, 1860. doi: 10.3390/s16111860

Beale, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J., and Elston, D. A. (2010). Regression analysis of spatial data. *Ecol. Lett.* 13, 246–264. doi: 10.1111/j.1461-0248.2009.01422.x

Bernal-Vasquez, A.-M., Möhring, J., Schmidt, M., Schönleben, M., Schön, C.-C., and Piepho, H.-P. (2014). The importance of phenotypic data analysis for genomic prediction - a case study comparing different spatial models in rye. *BMC Genomics* 15, 646. doi: 10.1186/1471-2164-15-646

Bini, L. M., Diniz-Filho, J. A. F., Rangel, T. F. L. V. B., Akre, T. S. B., Albaladejo, R. G., Albuquerque, F. S., et al. (2009). Coefficient shifts in geographical ecology: an empirical evaluation of spatial and non-spatial regression. *Ecography* 32, 193–204. doi: 10.1111/j.1600-0587.2009.05717.x

Bivand, R., and Piras, G. (2015). Comparing Implementations of Estimation Methods for Spatial Econometrics. *J. Stat. Softw.* 63, 1–36. doi: 10.18637/jss.v063.i18

Brenning, A. (2005). Spatial prediction models for landslide hazards: Review, comparison and evaluation. *Natural Hazards Earth Syst. Sci.* 5, 853–862. doi: 10.5194/nhess-5-853-2005

Cai, L., Bhattacharjee, A., Calantone, R., and Maiti, T. (2019). Variable selection with spatially autoregressive errors: a generalized moments LASSO estimator. *Sankhya B* 81, 146–200. doi: 10.1007/s13571-018-0176-z

Cressie, N. (1993). *Statistics for Spatial Data, Chapter 1*. New York, NY: John Wiley & Sons, Ltd.

Cruz, D. F., Meyer, S. D., Ampe, J., Sprenger, H., Herman, D., Hautegem, T. V., et al. (2020). Using single-plant-omics in the field to link maize genes to functions and phenotypes. *Mol. Syst. Biol.* 16, e9667. doi: 10.15252/msb.20209667

Day, B., Bateman, I., and Lake, I. (2004). *Ommitted locational variates in hedonic analysis: A semiparametric approach using spatial statistics*. Technical Report 04–04, Norwich.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-Based Geostatistics. *J. R. Stat. Soc. Ser. C Appl. Stat.* 47, 299–350. doi: 10.1111/1467-9876.00113

Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., et al. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* 30, 609–628. doi: 10.1111/j.2007.0906-7590.05171.x

Elias, A. A., Rabbi, I., Kulakow, P., and Jannink, J.-L. (2018). Improving genomic prediction in cassava field experiments using spatial analysis. *G3* 8, 53–62. doi: 10.1534/g3.117.300323

Fan, C., Rey, S. J., and Myint, S. W. (2017). Spatially filtered ridge regression (SFRR): A regression framework to understanding impacts of land cover patterns on urban climate. *Trans. GIS* 21, 862–879. doi: 10.1111/tgis.12240

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22. doi: 10.18637/jss.v033.i01

Fu, Z., Jiang, J., Gao, Y., Krienke, B., Wang, M., Zhong, K., et al. (2020). Wheat growth monitoring and yield estimation based on multi-rotor unmanned aerial vehicle. *Remote Sens.* 12, 508. doi: 10.3390/rs12030508

Ghosal, I., and Kormaksson, M. (2019). Plsmselect: Linear and Smooth Predictor Modelling with Penalisation and Variable Selection. *R PACKAGE Version 0.2.0*.

Haase, P. (1995). Spatial pattern analysis in ecology based on ripley's K-function: introduction and methods of edge correction. *J. Vegetat. Sci.* 6, 575–582. doi: 10.2307/3236356

Harisena, N. V., Groen, T. A., Toxopeus, A. G., and Naimi, B. (2021). When is variable importance estimation in species distribution modelling affected by spatial correlation? *Ecography* 44, 778–788. doi: 10.1111/ecog.05534

Harkel, J., Bartholomeus, H., and Kooistra, L. (2020). Biomass and crop height estimation of different crops using UAV-based lidar. *Remote Sens.* 12, 17. doi: 10.3390/rs12010017

Hodges, J. S., and Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *Am. Stat.* 64, 325–334. doi: 10.1198/tast.2010.10052

Jia, J., and Rohe, K. (2015). Preconditioning the Lasso for sign consistency. *Electron. J. Stat.* 9, 1150–1172. doi: 10.1214/15-EJS1029

Kissling, W. D., and Carl, G. (2008). Spatial autocorrelation and the selection of simultaneous autoregressive models. *Glob. Ecol. Biogeogr.* 17, 59–71. doi: 10.1111/j.1466-8238.2007.00334.x

Lado, B., Matus, I., Rodríguez, A., Inostroza, L., Poland, J., Belzile, F., et al. (2013). Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. *G3* 3, 2105-2114. doi: 10.1534/g3.113.007807

Langton, S. D. (1990). Avoiding edge effects in agroforestry experiments; the use of neighbour-balanced designs and guard areas. *Agroforestry Syst.* 12, 173–185. doi: 10.1007/BF00123472

Lee, W. J., Truong, H. A., Trinh, C. S., Kim, J. H., Lee, S., Hong, S.-W., et al. (2020). Nitrogen response deficiency 1-mediated chl1 induction contributes to optimized growth performance during altered nitrate availability in arabidopsis. *Plant J.* 104, 1382–1398. doi: 10.1111/tpj.15007

Lennon, J. J. (2000). Red-Shifts and Red Herrings in Geographical Ecology. *Ecography* 23, 101–113. doi: 10.1111/j.1600-0587.2000.tb00265.x

Liebhold, G. (2002). Integrating the statistical analysis of spatial data in ecology. *Ecography* 25, 553–557. doi: 10.1034/j.1600-0587.2002.250505.x

Ludwig, G., Zhu, J., Reyes, P., Chen, C.-S., and Conley, S. P. (2020). On spline-based approaches to spatial linear regression for geostatistical data. *Environ. Ecol. Stat.* 27, 175–202. doi: 10.1007/s10651-020-00441-9

Mao, X., Dutta, S., Wong, R. K. W., and Nettleton, D. (2020). Adjusting for spatial effects in genomic prediction. *J. Agric. Biol. Environ. Stat.* 25, 699–718. doi: 10.1007/s13253-020-00396-1

Meyer, H., Reudenbach, C., Wöllauer, S., and Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications–Moving from data reproduction to spatial prediction. *Ecol. Modell.* 411, 108815. doi: 10.1016/j.ecolmodel.2019.108815

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika* 37, 17–23. doi: 10.1093/biomet/37.1-2.17

Murakami, D. (2021). Spmoran: Moran Eigenvector-Based Scalable Spatial Additive Mixed Models. *R Package Version 0.2.1*.

Murakami, D., and Griffith, D. A. (2015). Random effects specifications in eigenvector spatial filtering: a simulation study. *J. Geogr. Syst.* 17, 311–331. doi: 10.1007/s10109-015-0213-7

Mylona, K., and Goos, P. (2021). Penalized generalized least squares for model selection under restricted randomization. *CiteSeerX [Preprint].*

Available online at: https://www.newton.ac.uk/documents/preprints/?search=Penalized+generalized+least+squares+

Nobre, W. S., Schmidt, A. M., and Pereira, J. B. M. (2020). On the effects of spatial confounding in hierarchical models. *Int. Statist. Rev*. 89, 302–322. doi: 10.1111/insr.12407

Paciorek, C. J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Stat. Sci*. 25, 107–125. doi: 10.1214/10-STS326

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and Team, R. C. (2021). Nlme: Linear and Nonlinear Mixed Effects Models. *R Package Version 3.1–152*.

Pinheiro, J. C., and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS, Vol. 100*. New York, NY: Springer.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., et al. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nat. Commun*. 11, 4540. doi: 10.1038/s41467-020-18321-y

Pohjankukka, J., Pahikkala, T., Nevalainen, P., and Heikkonen, J. (2017). Estimating the prediction performance of spatial models via spatial k-fold cross validation. *Int. J. Geogr. Inf. Sci*. 31, 2001–2019. doi,: 10.1080/13658816.2017.1346255

R., Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rabinowicz, A., and Rosset, S. (2020). Cross-validation for correlated data. *J. Am. Stat. Assoc*. doi: 10.1080/01621459.2020.1801451

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40, 913–929. doi: 10.1111/ecog.02881

Rocha, A. D., Groen, T. A., and Skidmore, A. K. (2019). Spatially-explicit modelling with support of hyperspectral data can improve prediction of plant traits. *Remote Sens. Environ*. 231, 111200. doi: 10.1016/j.rse.2019.05.019

Rocha, A. D., Groen, T. A., Skidmore, A. K., Darvishzadeh, R., and Willemen, L. (2018). Machine learning using hyperspectral data inaccurately predicts plant traits under spatial dependency. *Remote Sens*. 10, 1263. doi: 10.3390/rs10081263

Rocha, A. D., Groen, T. A., Skidmore, A. K., and Willemen, L. (2021). Role of sampling design when predicting spatially dependent ecological data with remote sensing. *IEEE Trans. Geosci. Remote Sens*. 59, 663–674. doi: 10.1109/TGRS.2020.2989216

Rodríguez-Álvarez, M. X., Boer, M. P., van Eeuwijk, F. A., and Eilers, P. H. C. (2018). Correcting for spatial heterogeneity in plant breeding experiments with P-splines. *Spat. Stat*. 23, 52–71. doi: 10.1016/j.spasta.2017.10.003

Romani, M., Borghi, B., Alberici, R., Delogu, G., Hesselbach, J., and Salamini, F. (1993). Intergenotypic competition and border effect in bread wheat and barley. *Euphytica* 69, 19–31. doi: 10.1007/BF00021722

Sarker, A., and Singh, M. (2015). Improving breeding efficiency through application of appropriate experimental designs and analysis models: a case of lentil (Lens culinaris Medikus subsp. Culinaris) yield trials. *Field Crops Res*. 179, 26–34. doi: 10.1016/j.fcr.2015.04.007

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., and Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecol. Modell*. 406:109–120. doi: 10.1016/j.ecolmodel.2019.06.002

Selle, M. L., Steinsland, I., Hickey, J. M., and Gorjanc, G. (2019). Flexible modelling of spatial variation in agricultural field trials with the R package INLA. *Theor. Appl. Genet*. 132, 3277–3293. doi: 10.1007/s00122-019-03424-y

Seya, H., Murakami, D., Tsutsumi, M., and Yamagata, Y. (2015). Application of LASSO to the eigenvector selection problem in eigenvector-based spatial filtering. *Geogr. Anal*. 47, 284–299. doi: 10.1111/gean.12054

Shijun, D., and Benzao, T. (2006). Generalized penalized least squares and its statistical characteristics. *Geospatial Inf. Sci*. 9, 255–259. doi: 10.1007/BF02826736

Singh, M., Malhotra, R. S., Ceccarelli, S., Sarker, A., Grando, S., and Erskine, W. (2003). Spatial variability models to improve dryland field trials. *Exp. Agr*. 39, 151–160. doi: 10.1017/S0014479702001175

Tang, Z., Parajuli, A., Chen, C. J., Hu, Y., Revolinski, S., Medina, C. A., et al. (2021). Validation of UAV-based alfalfa biomass predictability using photogrammetry with fully automatic plot segmentation. *Sci. Rep*. 11, 3336. doi: 10.1038/s41598-021-82797-x

Tiefelsdorf, M., and Griffith, D. A. (2007). Semiparametric filtering of spatial autocorrelation: the eigenvector approach. *Environ. Plann. A* 39, 1193–1221. doi: 10.1068/a37378

Verdooren, L. R. (2020). History of the statistical design of agricultural experiments. *J. Agric. Biol. Environ. Stat*. 25, 457–486. doi: 10.1007/s13253-020-00394-3

Wang, H., and Zhu, J. (2009). Variable selection in spatial regression via penalized least squares. *Can. J. Stat*. 37, 607–624. doi: 10.1002/cjs.10032

Wójtowicz, M., Wójtowicz, A., and Piekarczyk, J. (2016). Application of remote sensing methods in agriculture. *Commun. Biometry Crop Sci*. 11, 31–50. doi: 10.20546/ijcmas.2019.801.238

Wood, S. N. (2003). Thin-plate regression splines. *J. R. Stat. Soc.*) 65, 95–114. doi: 10.1111/1467-9868.00374

Yang, M., and Yuan, K.-H. (2019). Optimizing ridge generalized least squares for structural equation modeling. *Struct. Equ. Model*. 26, 24–38. doi: 10.1080/10705511.2018.1479853

Yoon, Y. J., Park, C., and Lee, T. (2013). Penalized regression models with autoregressive error terms. *J. Stat. Comput. Simul*. 83, 1756–1772. doi: 10.1080/00949655.2012.669383

Yoosefzadeh-Najafabadi, M., Earl, H. J., Tulpan, D., Sulik, J., and Eskandari, M. (2021). Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean. *Front. Plant Sci*. 11, 2169. doi: 10.3389/fpls.2020.624273

Yue, J., Yang, G., Li, C., Li, Z., Wang, Y., Feng, H., et al. (2017). Estimation of winter wheat above-ground biomass using unmanned aerial vehicle-based snapshot hyperspectral sensor and crop height improved models. *Remote Sens*. 9, 708. doi: 10.3390/rs9070708

Zhang, J., Virk, S., Porter, W., Kenworthy, K., Sullivan, D., and Schwartz, B. (2019). Applications of unmanned aerial vehicle based imagery in turfgrass field trials. *Front. Plant Sci*. 10, 279. doi: 10.3389/fpls.2019.00279

Zhou, X., Kono, Y., Win, A., Matsui, T., and Tanaka, T. S. T. (2021). Predicting within-field variability in grain yield and protein content of winter wheat using UAV-based multispectral imagery and machine learning approaches. *Plant Prod Sci*. 24, 137–151. doi: 10.1080/1343943X.2020.1819165