arXiv:2402.00183v1 [stat.ME] 31 Jan 2024

A review of regularised estimation methods and cross-validation in spatiotemporal statistics

Philipp Otto*
University of Glasgow, School of Mathematics and Statistics, United Kingdom
Alessandro Fassò
University of Bergamo, Italy
Paolo Maranzano
University of Milano-Bicocca, Fondazione Eni Enrico Mattei (FEEM), Italy

February 2, 2024

Abstract

This review article focuses on regularised estimation procedures applicable to geostatistical and spatial econometric models. These methods are particularly relevant in the case of big geospatial data for dimensionality reduction or model selection. To structure the review, we initially consider the most general case of multivariate spatiotemporal processes (i.e., $g > 1$ dimensions of the spatial domain, a one-dimensional temporal domain, and $q \geq 1$ random variables). Then, the idea of regularised/penalised estimation procedures and different choices of shrinkage targets are discussed. Finally, guided by the elements of a mixed-effects model, which allows for a variety of spatiotemporal models, we show different regularisation procedures and how they can be used for the analysis of georeferenced data, e.g. for selection of relevant regressors, dimensionality reduction of the covariance matrices, detection of conditionally independent locations, or the estimation of a full spatial interaction matrix.

Keywords: Big data, geostatistics, regularised estimation, review, spatiotemporal statistics

*Corresponding author: philipp.otto@glasgow.ac.uk

1 Introduction

In the era of big geospatial data, the analysis of intricate spatial and spatiotemporal processes has become both increasingly vital and challenging. This is attributed to the growing volume and finer resolution of geo-referenced data, including remotely sensed and LiDAR data, as well as the increasing diversity of data types due to automated collection and processing, such as social network data and image data. These advancements enable the examination of more intricate relationships and allow for the detection of weaker dependencies or variations in the data. To statistically investigate such interactions, geostatistical and spatial econometric models offer a robust framework for comprehending the fundamental structures of these processes. Additionally, they provide insights into spatial dependencies, temporal dynamics, and the interactions among multiple random variables. An essential advantage of these statistical models lies in their interpretability, unlike deep learning models, which are often viewed as black-box models unless explicitly designed for interpretability. However, as the dimensionality of the spatial domain increases and datasets become larger and more complex, conventional modelling approaches frequently encounter challenges related to computational complexity and the appropriate selection of influential variables.

Regularised estimation procedures offer effective solutions for these challenges from multiple perspectives. For instance, they can be used for model selection and dimensionality reduction but also to reveal spatial dependence structures going beyond geographical proximity. This article will provide a comprehensive review of regularisation techniques, including shrinkage and penalisation methods, used for statistical modelling of geospatial data. Thereby, our focus will be on geostatistical models (Section 3) and spatial econometric models (Section 4). Since the degree of regularisation is typically chosen based on predictive accuracy, we also review cross-validation techniques which are suitable under spatiotemporal dependence (Section 5). Finally, Section 6 concludes the article and discusses some direction for future research.

2 General framework

Geo-referenced q-variate data across space and time can be considered as a set of random variables $\{Y_t(\boldsymbol{s}) \in \mathbb{R}^q : \boldsymbol{s} \in D_{\boldsymbol{s}}, t \in D_t\}$ that are observed with a certain order in time and space. The spatial domain is denoted by $D_{\boldsymbol{s}} \subseteq \mathbb{R}^g$, which is a subspace with positive volume of

the g-dimensional real space with $g \geq 2$ (see Cressie and Wikle, 2015). By contrast, the time domain is one-dimensional, i.e., $D_t \subseteq \mathbb{R}$, typically assumed to be a subset of integers $D_t \subseteq \mathbb{Z}$, implying that the data are equidistantly observed in time. For D_t being a singleton $\{T\}$, i.e., one single time point T, the process collapses to a purely spatial process without a temporal dimension. Moreover, if $D_{\boldsymbol{s}}$ is a singleton of one location only, we get a q-variate time series process. There are different sources of (statistical) dependence for such geo-referenced data:

1. Temporal proximity between two observations typically causes correlations. That is, the closer two observations are in the temporal domain D_t, the higher they are correlated, e.g., the current outside temperature will be similar to the temperature one hour ago but less similar to the temperature 12 hours ago.

2. Geographical proximity in $D_{\boldsymbol{s}}$ induces a similar dependence, known as spatial dependence. As claimed by Tobler's first law of Geography, two observations are more similar if they are close to each other. That is, the current outside temperature in a city is similar to that in the neighbouring cities but less similar if you move away, especially in the North-South direction.

3. The q variables could be cross-correlated. Otherwise, if the q variables are (statistically) independent, one would rather prefer to model them independently using q univariate models. Moreover, network structures could describe the interrelations between these variables. In this case, network or graphical models for spatiotemporal data are suitable (see, e.g., Dey et al., 2021; Zapata et al., 2022, for graphical Gaussian processes for multivariate spatial data). However, graphical models are beyond the scope of this paper.

There are two (competing) approaches in spatial and spatiotemporal statistics to account for spatial dependence. Firstly, the dependence can be modelled using spatially dependent processes, where a suitable covariance function defines the entries of the covariance matrix. This approach is commonly known as geostatistics. Secondly, the dependent/outcome variable can be explicitly correlated with the nearby observations. Thereby, the local neighbourhood is defined by a suitable weight matrix $\mathbf{W}$, which is weighting all other observations, such that the product of $\mathbf{W}$ and the dependent variable is the weighted average of the adjacent observations. This approach is usually called the spatial econometrics approach, even though it is not tied to economic applications. Thus, we prefer to refer to this approach as spatial autoregression.

Below, we will provide an overview of regularised estimation procedures for both approaches, starting with geostatistical models in Section 3 and followed by spatial autoregression in Section 4.

3 Geostatistical models

Generally, there are different sources for increasing computational complexity. Firstly, if the number of spatial locations n increases, the covariance matrices of the process increase quadratically in n. It is worth noting that an increasing length of the time series is typically less critical because there is a causal relation in time (i.e., only past observations can influence future observations), such that spatiotemporal processes can be modelled conditionally on the past realisations. Secondly, if the number of variables q increases, we observe the same quadratic increase in the covariances between all variables. While the spatial correlation typically decreases with an increasing distance between the observations, which often allows for a sparse representation of the covariance matrices; the cross-correlation matrices are often dense.

Consider a mixed-effects spatiotemporal model, that is,

$$Y_t(\boldsymbol{s}) = \mu_t(\boldsymbol{s}) + \omega_t(\boldsymbol{s}) + \varepsilon_t(\boldsymbol{s}), \tag{1}$$

where $\mu_t(\boldsymbol{s})$ is the fixed-effects model, $\omega_t(\boldsymbol{s})$ is the random-effects model, and $\varepsilon_t(\boldsymbol{s})$ is the modelling error at time point t and location $\boldsymbol{s}$. While the fixed-effects term typically models the mean behaviour and the influence of exogenous variables, the random-effects term accounts for any additional dependence in the data, e.g., temporal and spatial dependence, heterogeneity, or cross-correlation between the variables. Now, let us distinguish between regularised estimation procedures for the parameters of $\mu_t(\boldsymbol{s})$ and $\omega_t(\boldsymbol{s})$.

Penalised estimation procedures can be used to shrink certain model parameters towards a pre-specified target. The most often applied shrinkage target is zero. In this case, the parameters are shrunk towards zero, which means that the parameter is excluded from the model if the estimated coefficient is equal to the zero target. Thus, these methods are suitable for model selection. When applied to the fixed effects model, it can be used for automatically selecting relevant regressors. In the following Section 3.1, we will focus on this in more detail. Moreover, a shrinkage target of zero can be used to introduce zeros in the spatial covariance matrix or the precision matrix, which would indicate the conditional independence between

these observations. This will be the subject of Section 3.2.

3.1 Regularisation for the large-scale variation

Linear regression models, whether referring to spatiotemporal or unstructured data, can be estimated by various techniques, including Likelihood maximisation and least squares. However, in high-dimensional settings, traditional methods for regression models might not be directly applied (Nandy et al., 2017). In order to reduce the model's complexity due to the fixed effect component $\mu_t(\boldsymbol{s})$, several methods for spatial and spatiotemporal settings have been proposed. Particularly, we refer to the task of selecting the most relevant predictors among a large set of candidates.

The state-of-the-art literature inherent to the statistical (algorithmic) variable selection task pivots around the Least Absolute Shrinkage and Selection Operator (LASSO) approach introduced by Tibshirani (1996). LASSO is a penalised version of ordinary least squares that uses a LASSO-type l_1 penalisation to shrink irrelevant parameters to zero. Under mild regularity conditions, including uncorrelated observations, LASSO ensures consistent parameter estimation (see Section 2.4.2 of Bühlmann and Van De Geer, 2011) and model selection consistency (Bickel et al., 2009; Belloni and Chernozhukov, 2013), i.e., LASSO owns the oracle property. Among others, one of the most relevant features of LASSO is that it admits solutions to the minimisation task even when the number of parameters is greater than the number of available observations. Also, LASSO penalty (and extensions) can be used either considering least squares and penalised likelihood problems (Fan and Li, 2001).

In the field of spatial and spatiotemporal regression, the number of observations can be very large, and the likelihood computation, even for Gaussian models, can require a very high computational effort (Stein et al., 2004a; Stein, 2014). Thus, when it comes to maximum likelihood estimation of space-time models, a major pathway is to rely on penalised maximum likelihood estimators (PMLE) of the parameters aiming at approximating the true likelihood function (Fan and Li, 2001; Fan and Peng, 2004; Zou and Li, 2008). However, asymptotic properties of approximate PMLEs rely on the asymptotic distribution of the initial estimators (e.g., ML estimates) used in the optimisation algorithm (Liu, 2017). To further improve computational efficiency, Chu et al. (2011) proposed a geostatistical version of PMLE in which the penalised function is approximated using one-step sparse estimator (Zou and Li, 2008) and

covariance tapering (Furrer et al., 2006a).

It is worth noting that real-world geostatistical applications are prone to cross-correlated regressors due to their dependent (spatially and temporally) structure. Zhao and Yu (2006) point out that when predictors are correlated, the classic LASSO algorithm does not provide selection-consistent estimates. Furthermore, when cross-correlation is detected, also group-LASSO estimators, which assume orthonormal data within each group, perform poorly in selecting the relevant predictors (Simon and Tibshirani, 2012). A straightforward solution to this issue is provided by the adaptive LASSO penalty, which leads to selection-consistent estimators even in the presence of cross-correlated covariates (see Zou, 2006) and Zou and Li (2008). Among others, Wen et al. (2018) use a spatial autoregressive model with adaptive LASSO penalty to detect relevant autoregressive parameters in genetic studies. Zhu et al. (2010) implemented an iterative penalised likelihood estimator with adaptive LASSO penalty to select predictors and neighbourhood structure in conditional autoregressive and simultaneous autoregressive models with spatially correlated error terms. Later on, Reyes et al. (2012a) proposed an adaptive LASSO algorithm for the case of linear regression models with spatiotemporal neighbourhood structures. Liu (2017) extended the previous algorithms, allowing for spatial correlation to be captured by either the spatial lag terms or spatial errors or both through a SARAR model. Also, their penalised estimates are obtained via least squares approximation to account for possible non-concavity of the likelihood function. Other examples of penalised likelihood for spatiotemporal data are in Al-Sulami et al. (2019), in which an adaptive LASSO method is proposed to simultaneously identify and estimate spatiotemporal lag interactions in the context of a data-driven semiparametric nonlinear model. Similarly, Liu (2022) developed an adaptive LASSO variable selection method for semiparametric spatial autoregressive panel models with random effects. The estimation is performed by maximising the concentrated profile likelihood function by means of a non-linear optimisation algorithm. Eventually, Chang et al. (2010) and Hsu et al. (2012) additionally reduced the model's complexity by combining covariance tapering and PMLE for spatial and spatiotemporal settings, respectively.

In addition to penalised likelihood methods, other estimation techniques have been developed for LASSO and its extensions in the geostatistical domain. For example, penalised least squares algorithms have been extended to the case of linear spatial models (Wang and Zhu, 2009), spatial autoregressive models (Cai and Maiti, 2020), regression models with spatially dependent data (Huang et al., 2010) and conditional autoregressive models (Gonella et al., 2022).

Also, when considering additive spatial models with potential non-linear effects, weighted versions of penalised least squares can be applied (Nandy et al., 2017). Cai et al. (2019) proposed a generalised moments LASSO, which combines LASSO with GMM estimator, to perform variable selection for spatial error models with spatially autoregressive errors. Huang et al. (2010) propose using LASSO to simultaneously select relevant predictors, choose neighbourhoods, and estimate parameters for spatial regression with GIS layers to predict responses in unsampled sites. Safikhani et al. (2020) considered LASSO methods for generalised spatiotemporal autoregressive models. The estimators are obtained by a modified version of the penalised least squares that accommodates hierarchical group LASSO-type penalties.Chernozhukov et al. (2021) combine least squares LASSO and bootstrap procedures to get estimates and inference for systems of high-dimensional regression equations characterised by temporal and cross-sectional dependences in covariates and error processes. Cao et al. (2022) proposed a penalised estimation procedure for Gaussian Processes regressions where the likelihood and the first two derivatives are approximated by means of a scaled Vecchia approximation (Vecchia, 2018). Eventually, several application-oriented papers combine classic LASSO approaches and geostatistical models in multi-step procedures (e.g., Fassò et al., 2022; Ye et al., 2011; Pejović et al., 2018).

Penalised methods are also commonly applied in the context of functional data analysis, especially involving penalised splines (see Silverman and Ramsay, 2002). These methods usually regularise the smoothness of the estimated functions by penalising the integrated second derivatives. In this way, many basis functions can be used, thus avoiding the typical overfit resulting from unpenalised estimation methods. Several authors have attempted to contribute by proposing LASSO-like penalised methods for selecting relevant functional (group) predictors (Pannu and Billor, 2017) or to identify regions where the coefficient function is zero and to smoothly estimate non-zero values of the coefficient function (Centofanti et al., 2022). Basis expansions and low-rank representations (Wood, 2017) are widely used tools in geostatistics for the spatiotemporal interpolation of environmental phenomena (see Hofierka et al., 2002; Xiao et al., 2016; Chang et al., 2010, for group-LASSO approaches in this context). For instance, in Maranzano et al. (2023), the authors make use of a PML estimator with adaptive LASSO penalty to select relevant functional covariates or their statistically relevant regions using the hidden dynamic geostatistical model. Eventually, Hsu et al. (2012) deal with semiparametric models for non-stationary spatiotemporal data in which a penalised least square with group-LASSO penalty is used to identify local spatial-temporal dependence features deviated from

the main stationary structure.

3.2 Regularisation for geospatial interactions

Second, suppose that the random effects follow a q-variate Gaussian process, i.e.,

$$\{\omega_t(\boldsymbol{s}) : \boldsymbol{s} \in D_{\boldsymbol{s}}, t \in D_t\} \quad \sim \quad N_p(0, C_\theta(\boldsymbol{s} - \boldsymbol{s}', t - t')), \tag{2}$$

where C_θ is a covariance function that depends on the difference between any two locations $\boldsymbol{s}$ and $\boldsymbol{s}'$ and two arbitrary time points t and t'. Moreover, if $q > 1$, C_θ gives a $q \times q$ covariance matrix for cross-variable correlation. This covariance function can depend on further unknown parameters θ, which have to be estimated. Here, in this case, the covariance function is called separable (Cressie and Wikle, 2015). Moreover, if C_θ is only a function of $||\boldsymbol{s}-\boldsymbol{s}'||$ with $||\cdot||$ being a vector norm (i.e., if only the distance between $\boldsymbol{s}$ and $\boldsymbol{s}'$ matters, but their orientation towards each other), the covariance function is isotropic. Generally, it is desirable if the covariance matrix resulting from C_θ contains many zeros; that is, it is sparse. The traditional way to induce zeros in the covariance is known as covariance tapering (Furrer et al., 2006b). Loosely speaking, based on the geographical distance between the locations, a covariance of zero is assumed for observations whose distance is larger than a certain threshold (theoretical results on covariance tapering can be found in Stein 2013). This method can also be applied in likelihood-based estimation procedures (see Kaufman et al. 2008, and Furrer et al. 2016 for theoretical results). As mentioned above, regularised estimation procedures are tailor-made to induce zeros for certain parameters when the shrinkage target is chosen to be zero. Typically, we can find zeros if two observations across space/time are conditionally independent (i.e., independent when observing all other realisations). However, the conditional independence is encoded in the inverse covariance matrix or precision matrix. Thus, penalised methods can be applied to obtain sparse precision matrices. For univariate data ($q = 1$), Krock et al. (2021a) introduced a graphical LASSO procedure to induce zeros in the precision matrix of a spatial process. Krock et al. (2021b) extended this approach to be used in the multivariate case ($q > 1$).

Moreover, since the covariance matrix is a positive definite matrix by definition (i.e., for valid choices C_θ, see Gneiting 2002; Gneiting et al. 2010; Stein 2005; Nychka et al. 2002; Porcu et al. 2016 or Porcu et al. 2021 for a historical review of space-time covariance functions), both the covariance matrix and its inverse can be decomposed as $\boldsymbol{\Sigma} = \mathbf{P}\mathbf{P}'$, e.g., via Cholesky decomposition. The matrix $\mathbf{P}$ is called Cholesky factor. Stein et al. (2004b) proposed to approximate

the likelihood for large spatial data sets based on Vecchia approximations (Vecchia, 1988). The idea is to approximate the joint likelihood as a product of the conditional likelihoods. Now, sparse Cholesky factors have to be considered first by Schäfer et al. (2021) for spatial models. If the dimension of $\mathbf{P}$ is $N \times r$ with N being the total number of observations across space/time and $r \ll N$, the covariance matrix is low-rank. That is, the dependence structure reduces to a lower-dimensional space. In spatial and spatiotemporal statistics, low-rank covariance matrices have been considered first by Banerjee et al. (2008), Cressie and Johannesson (2008) (spatial fixed-rank kriging), and Cressie et al. (2010) (spatiotemporal fixed-rank filtering). Back to the subject of this review paper, Chang et al. (2010) proposed a penalised estimation procedure to identify the lower rank of the covariance matrix. Specifically, low-rank approximations aim to represent a spatiotemporal process as a linear combination of local basis functions, which are weighted by uncorrelated random-effect coefficients (see also the recent review by Cressie et al. 2022). Thus, regularised estimation procedures aim to select suitable local basis functions. From a practical perspective, many different local basis functions (e.g., on several grids with different resolutions) can be included, and the best basis functions are chosen automatically by estimating the model parameters. For the spatiotemporal case, Hsu et al. (2012) suggested penalised procedures to choose these local basis functions. Furthermore, Kang and Katzfuss (2021) considered sparse inverse Cholesky factors that are identified based on correlations.

3.3 Bayesian estimation procedures

Along with the frequentist paradigm, the literature pioneered penalised regression extensions following a Bayesian perspective. Bayesian estimation schemes have gained particular importance for spatiotemporal models, especially using integrated nested Laplace approximations (INLA) which makes them applicable also for large data sets (see Rue et al., 2017, for a review on INLA in the spatiotemporal context). Moreover, we refer the readers to the review paper by van Erp et al. (2019) for a comprehensive overview of the state-of-the-art literature on Bayesian penalised regression. The authors summarise that Bayesian penalisation techniques include the penalised in three alternative ways. The first way is called *Full Bayes* or *hierarchical Bayes* (Wolpert and Strauss, 1996) approach, which treats the penalty parameter λ as an unknown variable (i.e., a hyperparameter) whose prior distribution has to be specified. Such models specify prior distributions for all parameters and can be estimated in a single step. The prior

distribution is called the shrinkage prior (van Erp et al., 2019) and is usually a vague distribution, e.g., a half-Cauchy random variable (Polson and Scott, 2012). The prior acts on the coefficients in order to shrink small effects to zero while maintaining true large effects. Indeed, large values of λ result in smaller prior variation and thus more shrinkage of the coefficients towards zero.

The second way is named *Empirical Bayes* (van de Wiel et al., 2019) and intends the parameters as unknown constants. This approach differs from the first one because it involves a two-step process: first, estimating the penalty parameter λ from the observed data, and second, incorporating this empirical estimate into the model using an empirical Bayes prior distribution. Since the empirical approach does not specify any prior distribution for the hyperparameters, a sensitivity analysis of the results with respect to the distributions is not necessary.

The third approach is based on cross-validation (CV). In this case, there is no difference between the frequentist and Bayesian frameworks, as the goal is to select λ so that the model is as accurate as possible in predicting new values of the response variable. The topic of spatiotemporal CV will be extensively discussed in the following Section 5.

As mentioned above, the fully Bayesian approach requires defining an a priori distribution for the penalty term of each hyperparameter. Different distributions were proposed depending on the penalised regression type (e.g., LASSO, ridge, or elastic net). For example, in the case of ridge regression (Hastie, 2020), the ridge prior corresponds to a Normal centred on the origin (Hsiang, 1975), while in the case of LASSO, a Laplace distribution is employed (Park and Casella, 2008). For a Bayesian LASSO Gibbs sampler, the Laplace distribution can be represented as a scale mixture of Gaussians (with an exponential mixing density). It is worth noting that the ridge regression estimator can be viewed as the Bayesian posterior mean estimator of the coefficients when imposing a Gaussian prior on the regression parameter (van Wieringen, 2015). Further extensions can be found in van Erp et al. (2019), in which the authors compare several shrinkage priors from theoretical and application perspectives.

In addition to the penalised regression approach discussed above, the Bayesian framework includes other model selection techniques, such as the Zellner's g-prior (Zellner, 1986), and the *spike-and-slab prior* (Mitchell and Beauchamp, 1988). The former approach shrinks the regression coefficients toward zero through a global shrinkage scalar called g, which equally shrinks each coefficient (which can be reasonable if the coefficients are equivalent). The single-

g Zellner's prior has been extended in several ways, e.g. using a mixture of g priors (Liang et al., 2008), and multiple shrinkage factors as in Zhang et al. (2016). However, carefully choosing the constant g must be addressed to avoid excluding important variables (Lindley, 1957). Instead, according to the data, the spike-and-slab method assigns the regression coefficients to the zero-centred spike (i.e., shrinking toward zero) if they do not deviate substantially from zero. In contrast, if they differ significantly from zero, they will be assigned to the slab (i.e., the vague proper prior). Both methods were also adapted to spatio-temporal analysis. Refer, for example, to Lee et al. (2014) on the combination of Zellner's g prior and spatial Ising prior for selecting spatial covariates in spatial time series data.

The literature on Bayesian modelling for spatiotemporal data addresses the analysis through both a fully Bayesian hierarchical approach (Wikle et al., 1998) and an empirical Bayes approach (Fahrmeir et al., 2004). Fully Bayesian frameworks were extended to the case of variable selection in large spatiotemporal models in several ways. Katzfuss and Cressie (2012) proposed a Bayesian hierarchical spatiotemporal random effects where dimensionality reduction is achieved by means of spatiotemporal basis functions, whereas the prior induces sparsity and shrinkage on the first-order autoregressive parameters describing the temporal evolution of the basis-function coefficients. The described approach was inspired by the so-called *Minnesota prior* (Ingram and Whiteman, 1994; George et al., 2008). Such prior was initially developed in a time series context where the aim was to drop the autoregressive coefficients of VAR models by shrinking the posterior of the parameter matrix towards independent random walk models (the typical behaviour of stock prices in financial applications).

Possible alternatives to penalty methods for selecting linear predictors in space-time models are mixture model selection methods. This category includes Bayesian selection methods and Bayesian model averaging. The former evaluates the appropriateness of a model based on the estimated weight among a variety of models with alternative predictors; the latter, on the other hand, averages over several alternative models to find the posterior distribution of the parameters. Spatial dynamics is included in the algorithms by adding an intrinsic conditional autoregressive (ICAR) (Besag et al., 1991; Besag and Green, 1993) approach (Carroll et al., 2018), while temporal dynamics is incorporated via autoregressive processes (Lawson et al., 2017). Both methods have proven effective in disease mapping studies with spatial data (Carroll et al., 2018), spatial small area frameworks (Carroll et al., 2016a), as well as spatio-temporal disease mapping (Carroll et al., 2016b). Among others, relevant advantages include getting an

automatic final model fit.

4 Spatial autoregression

Since this approach requires the explicit definition of the neighbourhood structure via $\mathbf{W}$, the prediction at unknown locations is more complicated (it would require a distance-dependent functional relation of each weight), and the models are usually applied to panel data. That is, the observational sites are typically constant across time. Suppose that there are n different locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ and $\boldsymbol{Y}_t = (Y_t(\boldsymbol{s}_1), \ldots, Y_t(\boldsymbol{s}_n))'$, then a simple spatial autoregressive process (or simultaneous autoregressive process) without a temporal autoregressive dependence would be given by

$$\boldsymbol{Y}_t = \mathbf{X}_t\boldsymbol{\beta} + \rho\mathbf{W}\boldsymbol{Y}_t + \boldsymbol{\varepsilon}_t\,, \tag{3}$$

where $\mathbf{X}_t$ is a matrix of regressors, $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients, ρ is the spatial autoregressive parameter, and $\boldsymbol{\varepsilon}_t$ is the vector of model errors. For spatial autoregressive (SAR) and conditional autoregressive (CAR) models, Gonella et al. (2022) proposed a LASSO estimation procedure to select the relevant covariates in the regression term.

Assuming normal random errors with covariance matrix $\boldsymbol{\Sigma}_\varepsilon$, this approach can also be considered as a mixed-effects spatiotemporal model with

$$\begin{aligned}
\boldsymbol{\mu}_t = (\mu_t(\boldsymbol{s}_1), \ldots, \mu_t(\boldsymbol{s}_n))' &= (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}_t\boldsymbol{\beta}\,, \\
\boldsymbol{\varepsilon}_t = (\varepsilon_t(\boldsymbol{s}_1), \ldots, \varepsilon_t(\boldsymbol{s}_n))' &\sim N_n(\mathbf{0}, (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\Sigma}_\varepsilon(\mathbf{I} - \rho\mathbf{W}')^{-1})\,.
\end{aligned}$$

Alternatively, the spatial interactions can be modelled in the random effects term in the same manner, while the errors remain independent across space. Generally, there is a relation between geostatistical and spatial autoregression models, and both approaches are equivalent under certain conditions (Ver Hoef et al., 2018, Theorem 1).

The precision matrix of such spatial autoregressive models is given by

$$(\mathbf{I} - \rho\mathbf{W})'\boldsymbol{\Sigma}_\varepsilon^{-1}(\mathbf{I} - \rho\mathbf{W}) \tag{4}$$

showing the relation to the above-mentioned Cholesky decomposition of geostatistical models. Thereby, the spatial weight matrix $\mathbf{W}$ implies a certain (geographical) structure of the Cholesky factors. In this general framework, Zhu and Liu (2009) proposed a LASSO procedure to estimate

the precision matrix, exploiting the fact that geographically distant observations are likely to be conditionally independent (i.e., the precision matrix is a sparse matrix). In this way, the zero entries can be automatically identified.

Instead of the linear relation $\rho\mathbf{W}$, the spatial interactions can be modelled using a series of different weight structures, e.g., $\sum_{i=1}^{k} \rho_i \mathbf{W}_i$ with k different weight matrices $\mathbf{W}_i$. For instance, each weight matrix could only contain the weights for certain directions (northward, north-eastward, eastward dependence, etc.) to reveal directional processes (Merk and Otto, 2021). Moreover, penalised estimation procedures can be used to select the true weight matrix $\mathbf{W}$ from a series of alternative weights $\mathbf{W}_1, \ldots, \mathbf{W}_k$, as for the boosting procedure proposed by Kostov (2010, 2013) or the LASSO least-squares procedure proposed by Lam and Souza (2020). Reyes et al. (2012b) applied a spatiotemporal LASSO procedure to select weight matrices from a set of candidates with increasing spatial lag order, simultaneously with the temporal lags and spatiotemporal weight matrices. In other words, they constructed a 2-dimensional grid of the temporal and spatial lag orders and selected the relevant spatiotemporal interactions. In the context of traffic analysis, Haworth and Cheng (2014) applied a graphical LASSO procedure to select local neighbourhood structures analogously. Together with the autoregressive coefficients, Reyes et al. (2012b) also penalised the regressive parameters using a second penalty term. Similarly, Liu et al. (2018) suggested a LASSO procedure for selecting the regressors in the mean equation while allowing for spatial autoregressive structure in the model. Since ordinary least-squares procedures are inconsistent in the presence of spatial autoregressive dependence, they proposed a penalised quasi-maximum likelihood approach incorporating a LASSO penalty with a zero shrinkage target for the regression coefficients.

In addition to these approaches, there are several attempts to fully estimate the spatial weight matrix $\mathbf{W}$ using regularised procedures. Due to complex interactions and high flexibility, the main issue is to uniquely identify each weight (Manski 1993, and Gibbons and Overman 2012 for a critical review of spatial econometric procedures). That is, if one weight between region A and B, say w_{ab}, is misspecified, this can be compensated via further linkages through other locations, e.g., via w_{ac} and w_{cb}, and still lead to the same spatial covariance matrix. The same applies to the distinction between directed links between A and B and vice versa (i.e., w_{ab} and w_{ba}). To the best of our knowledge, Zhu and Liu (2009) and Bhattacharjee and Jensen-Butler (2013) first introduced the idea of estimating the full matrix $\mathbf{W}$, where they implied further structural constraints for identification. To be precise, they assumed a triangular weight matrix

or symmetric dependence structure. Another structural constraint, namely a block-diagonal structure, was considered in Lam and Souza (2016). Further, Ahrens and Bhattacharjee (2015) proposed a two-step LASSO procedure to estimate the spatial weight matrix in spatial autoregressive models. For spatial lag models, a regularised estimation procedure was introduced by Lam and Souza (2020). Under the assumption of locally constrained spatial dependence, Merk and Otto (2022) suggested an adaptive LASSO procedure based on cross-sectional resampling, which makes the estimation scalable for large datasets. For spatiotemporal data with unknown structural breaks in the mean, a constraint two-step LASSO procedure was introduced by Otto and Steinert (2022). Their method could estimate the spatial weight matrix together with all structural breaks and the positions of the change points.

5 Cross validation under spatiotemporal dependence

All regularised estimation procedures require the choice of the degree of regularisation via a so-called penalty parameter, often denoted by λ. If $\lambda = 0$, the models coincide with their unpenalised version, whereas the degree of penalisation, and thus, the shrinkage of the parameters towards the shrinkage target, increases with an increasing value of λ.

When a strictly positive λ is used, the estimated coefficients β_λ are biased but are more efficient, i.e., their variance is smaller. This is known as the bias-variance tradeoff. Additionally, opting for a λ value that is too small can result in overfitting for richly parametrised models, while selecting a value that is too large can lead to underfitting (see, e.g., Boonstra et al., 2015). The optimal penalty parameter is usually selected based on the model's in-sample or out-of-sample predictive performance. Alternative ways to select the penalty parameter are likelihood-based methods, where λ is interpreted as a variance component and the likelihood is maximised with respect to the couple $(\sigma_\varepsilon^2, \lambda)'$. Similarly, the penalty parameter can be simultaneously estimated with all parameters in a fully Bayesian approach, where typically vague half-Cauchy prior distribution is assumed for λ (van Erp et al., 2019). Alternatively, Otto and Steinert (2022) proposed to select the penalty parameter based on the distance between the sample and model spatial autocorrelation. Below, we will focus on the predominantly applied goodness-of-fit (GoF) criteria.

GoF criteria aim to maximise the model fit by tuning the penalty parameter λ. The

model fit could be assessed by the in-sample fit in terms of distance between the observed and predicted values or the out-of-sample fit, i.e., cross-validation prediction accuracies. The distance is typically evaluated by Root-Mean-Squared-Error (RMSE), Mean Absolute Error (MAE), or based on information criteria, such as Akaike's Information Criterion (Akaike, 1973, AIC) and the Bayesian Information Criterion (Schwarz, 1978, BIC). Most models have no closed-form solutions to determine the optimal value of λ, necessitating grid-search algorithms. Arlot and Celisse (2010) offers a thorough review of the significance of cross-validation (CV) in regression model selection and its statistical properties.

In general, a valid CV must satisfy three properties (Jiang and Wang, 2017):

1. randomness of partition
2. mutual independence of test errors
3. independence between the training set and test set.

Thus, depending on the data structure (e.g., cross-sectional, clustered, spatial, temporal data), the cross-validation schemes to be adopted may vary significantly. In the case of independent data, classical CV schemes, such as random k-fold, stratified k-fold (Zeng and Martinez, 2000; Ludwig et al., 2016) or Generalized CV (Boonstra et al., 2015) can be used. For instance, considering ridge regularisation, the GCV estimator can be used to efficiently estimate the penalty parameter λ even when the number of observations (thus, the degrees of freedom) is small or the number of parameters to be estimated exceeds the number of observations (Golub et al., 1979).

In the case of spatially/temporally dependent data, standard random k-fold cross-validation procedures cannot be applied. Schratz et al. (2019) showed that, while performances are over-estimated when no spatial information is included in the CV step, hyperparameter tuning of machine learning models appears to be less sensitive to the spatial structure, leading to similar results of non-spatial and spatial CV schemes. The problems arising from the occurrence of spatio-temporal autocorrelation in model assessment are manifold.

First, using random sampling in cross-validation results in test observations being collected from areas that are spatially close to the training observations (Schratz et al., 2019). Consequently, the evaluation of prediction performance tends to be overly optimistic because

the training and test datasets become correlated, largely due to the neglected underlying correlation structure, whether across space or time (Brenning, 2012; Meyer et al., 2019; Meyer and Pebesma, 2021; Ploton et al., 2020; Lezama Valdes et al., 2021). That is, if one leaves out a set of observations and estimates a spatiotemporal model with the remaining observations, information from the observations used for model estimation is used to predict the left-out observations (via the spatiotemporal interactions). Second, when the spatiotemporal structure is neglected, the estimated residuals will not be mutually independent, which is a critical assumption in many statistical models. As a result, it is often advisable to exclude complete data blocks across time and/or space (Roberts et al., 2017).

For stationary time series, cross-validation (CV) is applied by segmenting the time series into subsets. Two main strategies are commonly used: (1) k-fold blocked subsets CV, where each subset is the test set in turn, and the rest are for training and forecasting (Bergmeir and Benítez, 2012; Bergmeir et al., 2014, 2018), and (2) last-block CV or forward validation CV, where only the final block is the test set, and previous blocks are for training (Hjorth and Hjort, 1982). The key distinction between the two approaches lies in preserving temporal order. In the last-block strategy, the model is never tested on past data relative to the training data. For instance, an hv-block method can be used to maintain independence between training and test sets, excluding a window of h observations before and after the test set (Racine, 2000). Moreover, both strategies allow choosing among several combinations of forecast horizons and updating schemes for the two samples (Tashman, 2000, e.g., fixed-origin, rolling-origin and recalibration) to compute forecasting accuracy metrics (Bergmeir and Benítez, 2012). A graphical synthesis of the two strategies is reported in Figure 1. Further CV strategies for temporal data can be found in Jiang and Wang (2017) (Markov-CV and partitioned-CV), Cerqueira et al. (2020) (prequential approach), and Cerqueira et al. (2017) (Monte Carlo replications last-block approach). In the case of non-stationary time series, the previous CV schemes can be strongly misleading, as the unknown future may differ from the training sample, the test sample, or both. In this case, Bergmeir and Benítez (2012) suggest using a weighted overlapping approach in which the whole series is used in training and testing steps.

Spatial data share challenges regarding autocorrelation with time series, but spatial data involves at least two-dimensional coordinates, whereas time is one-dimensional (Roberts et al., 2017). To address these issues, spatial CV methods and spatial variable selection techniques, like recursive and forward spatial feature selection, can be employed (Meyer et al., 2018, 2019).

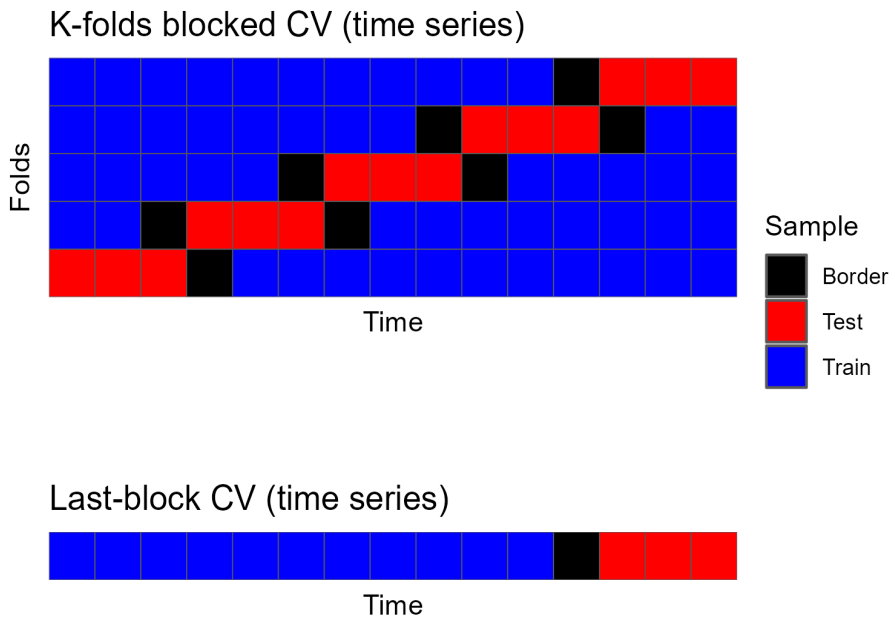

Figure 1: Temporal CV schemes. Blue blocks represent points from the time series in the training set, and red blocks represent points in the test set. Black blocks are neighbourhood observations to the test set and omitted from the training set.

In addition, Schratz et al. (2019) demonstrated that hyperparameter tuning in machine learning approaches is less sensitive to spatial structure, but they recommend spatial cross-validation to ensure unbiased predictive performance.

Moreover, Meyer and Pebesma (2021) introduced the "area of applicability" (AOA) concept, defining the model's valid geographical area based on training data. That is, the maximum distance (without outliers further apart than 1.5 interquartile ranges) in the covariate space of the training data defines the AOA in the prediction space. All predictions which are further apart than the maximum distance are marked as outside the AOA. The AOA will also account for the geographical distance if geographical coordinates are included in the covariates. In this way, the idea prevents predicting new geographic spaces with conditions that are very different from the training data, where the models can dramatically fail (Meyer and Pebesma, 2022).

The time-series block CV structure can be easily extended to spatial data by partitioning observations into spatial blocks. These blocks can be created either by dividing the entire space into cells for gridded data or by establishing spatial buffers around the data collection points for network or point data (Roberts et al., 2017). In Figure 2, we represent examples of spatial buffering both using grids and point data. In the latter, when the CV is performed by iteratively eliminating one location (and the neighbours within the buffer) at a time, we refer to spatial leave-one-location-out (SLOO) CV (Gasch et al., 2015; Meyer et al., 2018, 2019). As per **?**, SLOO yields a criterion similar to the AIC but without accounting for spatial autocorrelation, meaning it produces the same output as AIC-based model selection in this context. However, when spatially correlated variables are in the model, AIC may not select the right covariates, whereas SLOO performs better. Additionally, spatial blocking can be applied to point patterns by assigning each location to its corresponding training polygon.

When multiple locations are used to build the test set, we refer to spatial k-fold CV (Pohjankukka et al., 2017). To maintain independence between training and test sets, blocks bordering the test set (for spatial blocking) or locations within the buffer (for point data) are excluded from the training set (Milà et al., 2022, b-LLO proposed by). The most relevant problem in this situation is to determine what is the optimal buffer length that guarantees independence (Trachsel and Telford, 2016).

Possible proposals include (1) fitting a *prior* variogram to the raw data and using the resulting distance as block length (Bio et al., 2002); (2) estimating the autocorrelation range fit-

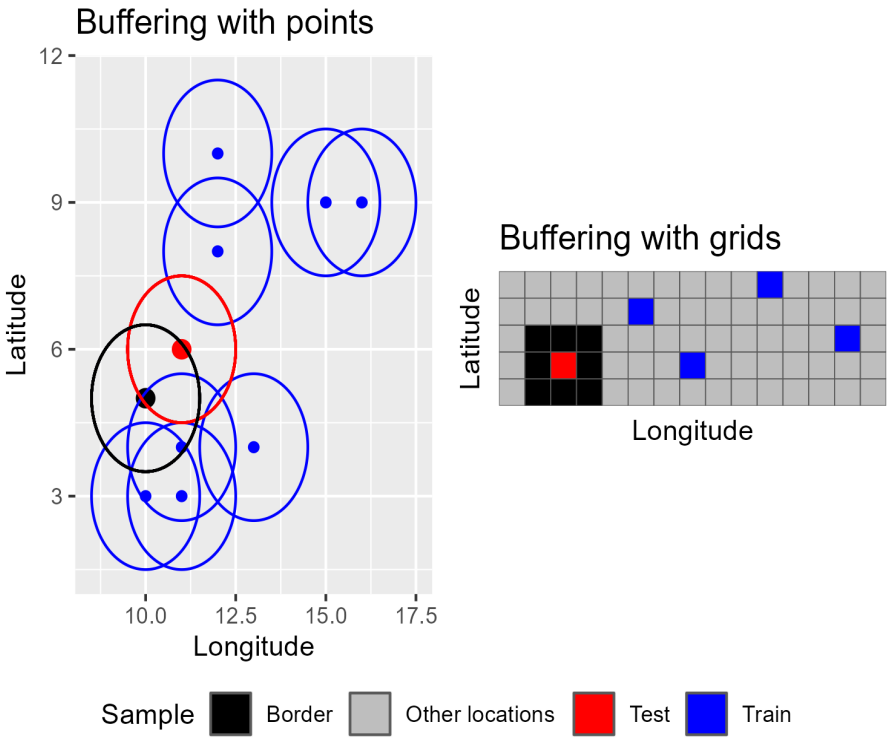

Figure 2: Spatial buffering for point data (left panel) and for gridded data (right panel).

ting a variogram (Brenning, 2005; Roberts et al., 2017; Milà et al., 2022) or a circular variogram (Telford and Birks, 2009) to the model residuals; (3) implementing a spatial independence test to find the minimum distance such that taken one point in the test set and one in the training set they are uncorrelated (Telford and Birks, 2005). However, as pointed out by Brenning (2022), residuals-based solutions for estimating the autocorrelation range are model-dependent. Thus, the necessity arises for model-agnostic validation tools to assess how predictive performance degrades with increasing prediction distances. The author introduces spatial prediction error profiles (SPEPs), which link the median prediction distances to the spatial prediction errors on the test set. This method can be used to understand (1) how the CV performance of the model decays for increasing distances from the training set and (2) how competing models perform in predicting values at large and small distances.

Alternative approaches to spatial blocking have been proposed in recent geostatistics literature. For instance, geographical partitioning provides a valid alternative via k-means algorithm (Brenning, 2012). Provided a fixed number of partitions k, the clustering algorithm partitions the spatial location into non-overlapping clusters based on geodesic distance. Unfortunately, the simple k-means algorithm does not permit controlling the number of points in

each partition, leading to potential heterogeneous partitions. To overcome this problem, Wang et al. (2021) proposed a heuristically modified k-means algorithm favouring partitions with similar elements. Also, see the inverse sampling-intensity weighting system by de Bruin et al. (2022) in which one can assign more weight to observations in sparsely sampled areas and less weight to observations in densely sampled areas to correct for estimation bias.

When spatio-temporal data are considered, block-based CV schemes are obtained as a combination of the previously cited spatial and temporal blocking strategies. Following Meyer et al. (2018), we refer to time-block partitioning as Leave-Time-Out (LTO), point-in-space partitioning as Leave-Location-Out (LLO), and space-time partitioning as Leave-Location-and-Time-Out (LLTO). The three strategies are also called *target-oriented* to contrast the classical *random* approach. Specifically, the LTO partitions the spatiotemporal observations along the time axis into blocks of time series common to all locations (i.e., it performs training by eliminating all time instants assigned to the test block for all spatial locations and iterates over the time blocks). LLO does the same, but iterates w.r.t. the spatial blocks and considers the whole time series of each location. LTLO iterates over both temporal and spatial blocks (i.e., at a given iteration, the algorithm eliminates all observations of a certain spatial block and, for all other blocks, eliminates a common temporal block). Note that LLO coincides with the spatial blocking presented above, whereas LTO coincides with the temporal k-fold block strategy. In Figure 3, we show a schematic example of the three CV schemes for spatio-temporal data.

A further extension of the LLO approach is the Nearest Neighbour Distance Matching (NNDM) LOO CV introduced by Milà et al. (2022). This variant compares the nearest neighbour distance distribution function between the test and training data in the CV process to the nearest neighbour distance distribution function between the target prediction and training points. In practice, this is an alternative method to b-LLO in which the neighbours to be excluded are defined not by distance from the point but by the mismatch between the two Nearest Neighbour Distance distributions.

Spatial and spatio-temporal resampling methods (CV and bootstrap) are implemented in many statistical software, such as R (R Core Team, 2021). Among the others, see the packages **sperrorest** (Brenning, 2012) (which implements distance-based K-means spatial partitioning); **blockCV** (Valavi et al., 2019), which implements block partitions and buffering for spatial data, as well as providing geostatistical tools for measuring spatial autocorrelation ranges in candidate covariates for model training and simplifying the choice of block and buffer sizes. It also offers

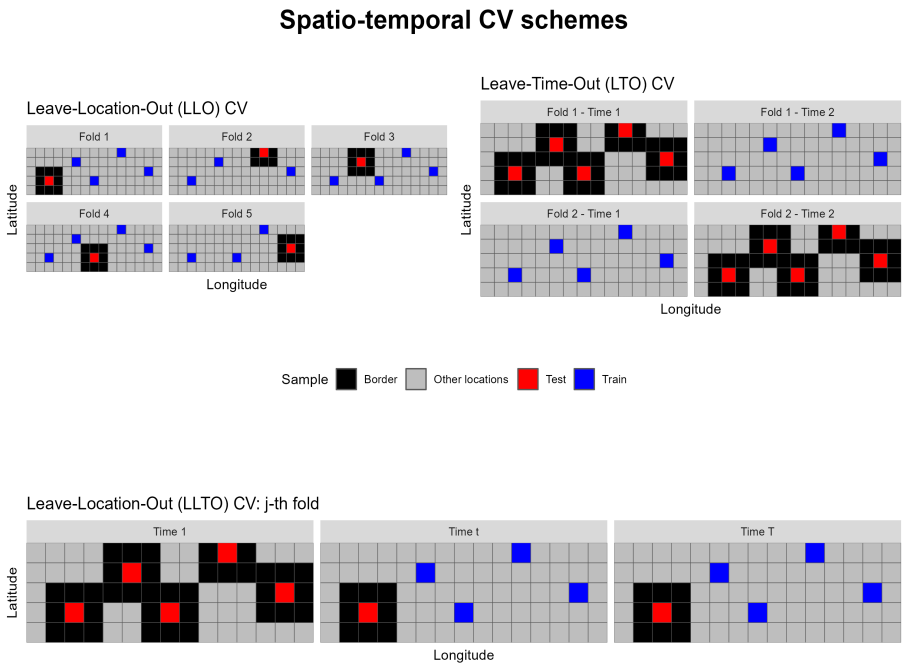

Figure 3: Spatio-temporal CV schemes. Blue blocks represent points used for training the model, while red blocks represent points in the test set. Black blocks are neighbourhood observations to the test set and omitted from the training set.

an interactive tool for visualising spatial blocks as a function of folds and block/buffer sizes; **CAST** (Meyer et al., 2022), which implements the NNDM CV scheme as well as performing spatial variable selection to select suitable predictor variables according to their contribution to the spatial model performance; and **mlr3spatiotempcv** (Schratz et al., 2021), a package that is part of the **mlr3** ecosystem (Lang et al., 2019) for the unified implementation of a wide range of ML models with feature and model selection tools and model evaluation capabilities. Specifically, **mlr3spatiotempcv** implements k-fold temporal and spatial blocking partitioning with and without buffering described in Meyer et al. (2018), and resumes the other partitioning techniques used in **sperrorest**, **blockCV**, **skmeans** (Zhao and Karypis, 2002, hierarchical agglomerative clustering algorithms by), and **CAST**.

6 Conclusion and directions for future research

In the field of spatial statistics and spatial econometrics, this comprehensive review has demonstrated the vital role of penalised methods in coping with the growing complexity of modern spatial data. With the increasing availability of geo-referenced data in various formats and types, the demand for adaptable, interpretable, and efficient modelling approaches becomes increasingly evident. This is where regularisation techniques step in, emerging as versatile tools for model selection, dimensionality reduction, and exploring spatial dependencies for classic statistical models. The advantage of interpretability sets them apart from the often enigmatic nature of deep learning models.

In our review paper, we have presented the landscape of different regularisation methods, from the nuances of shrinkage to the mechanisms of penalisation strategies. We have underlined their practicality and effectiveness in the statistical modelling of geospatial data. We also briefly looked at Bayesian regularised estimation methods. Since the regularised estimation methods require the choice of a penalty term, which is usually done by cross-validation, we have also summarised cross-validation methods that can be used in the case of spatiotemporal dependence.

Despite the already substantial literature, integrating and extending regularisation techniques into geostatistical modelling is a promising approach for the future. One notable avenue is the application of regularisation methods for estimating spatial covariance functions, offering

novel insights into spatial relationships. A promising approach is estimating spatial covariance functions using regularised splines, which allow an automated choice of basis functions by regularising the smoothness of the estimated function. However, the difficulty lies in ensuring that the covariance function is valid, i.e. that it generates positive-definite covariance matrices.

There is a need for a more extensive exploration of regularised methods in spatial econometrics, particularly in estimating weighting matrices. This matrix is usually assumed to be known, which is rarely the case in practice, but allows the results to be interpreted in a geographical sense. To enhance the interpretability of these estimated matrices, one avenue to consider is using traditional distance-based weighting matrices as shrinkage targets, thereby enabling a geographical interpretation. Generally, the quest for enhanced computational efficiency in the application of regularisation methods for large-scale geospatial applications remains a pertinent concern in both fields.

In summary, in a world where the diversity and volume of geospatial data continue to increase, this review paper is intended to provide guidance for understanding regularised methods in spatial and spatiotemporal statistics to advocate their use for geospatial analyses.

References

Ahrens, A. and Bhattacharjee, A. (2015). Two-step lasso estimation of the spatial weights matrix. *Econometrics*, 3(1):128–155.

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle*. Akadémiai Kiadó, Budapest, Hungary.

Al-Sulami, D., Jiang, Z., Lu, Z., and Zhu, J. (2019). On a semiparametric data-driven nonlinear model with penalized spatio-temporal lag interactions. *Journal of Time Series Analysis*, 40(3):327–342.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4(none):40–79, 40.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213.

Bergmeir, C., Costantini, M., and Benítez, J. M. (2014). On the usefulness of cross-validation for directional forecast evaluation. *Computational Statistics & Data Analysis*, 76:132–143.

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83.

Besag, J. and Green, P. J. (1993). Spatial statistics and bayesian computation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(1):25–37.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20.

Bhattacharjee, A. and Jensen-Butler, C. (2013). Estimation of the spatial weights matrix under structural constraints. *Regional Science and Urban Economics*, 43(4):617–634.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 28.

Bio, A. M. F., De Becker, P., De Bie, E., Huybrechts, W., and Wassen, M. (2002). Prediction of plant species distribution in lowland river valleys in belgium: modelling species response to site conditions. *Biodiversity & Conservation*, 11(12):2189–2216.

Boonstra, P. S., Mukherjee, B., and Taylor, J. M. (2015). A small-sample choice of the tuning parameter in ridge regression. *Statistica Sinica*, 25(3):1185.

Brenning, A. (2005). Spatial prediction models for landslide hazards: review, comparison and evaluation. *Nat. Hazards Earth Syst. Sci.*, 5(6):853–862.

Brenning, A. (2012). Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The r package sperrorest. In *2012 IEEE International Geoscience and Remote Sensing Symposium*, pages 5372–5375.

Brenning, A. (2022). Spatial machine-learning model diagnostics: a model-agnostic distance-based approach. *International Journal of Geographical Information Science*, pages 1–23.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Cai, L., Bhattacharjee, A., Calantone, R., and Maiti, T. (2019). Variable selection with spatially autoregressive errors: A generalized moments lasso estimator. *Sankhya B*, 81(1):146–200.

Cai, L. and Maiti, T. (2020). Variable selection and estimation for high-dimensional spatial autoregressive models. *Scandinavian Journal of Statistics*, 47(2):587–607.

Cao, J., Guinness, J., Genton, M. G., and Katzfuss, M. (2022). Scalable gaussian-process regression and variable selection using vecchia approximations. *J. Mach. Learn. Res.*, 23(1):Article 348.

Carroll, R., Lawson, A. B., Faes, C., Kirby, R. S., Aregay, M., and Watjou, K. (2016a). Bayesian model selection methods in modeling small area colon cancer incidence. *Annals of Epidemiology*, 26(1):43–49.

Carroll, R., Lawson, A. B., Faes, C., Kirby, R. S., Aregay, M., and Watjou, K. (2016b). Spatio-temporal bayesian model selection for disease mapping. *Environmetrics*, 27(8):466–478.

Carroll, R., Lawson, A. B., Faes, C., Kirby, R. S., Aregay, M., and Watjou, K. (2018). Spatially-dependent bayesian model selection for disease mapping. *Statistical Methods in Medical Research*, 27(1):250–268.

Centofanti, F., Fontana, M., Lepore, A., and Vantini, S. (2022). Smooth lasso estimator for the function-on-function linear regression model. *Computational Statistics & Data Analysis*, 176:107556.

Cerqueira, V., Torgo, L., and Mozetič, I. (2020). Evaluating time series forecasting models: an empirical study on performance estimation methods. *Machine Learning*, 109(11):1997–2028.

Cerqueira, V., Torgo, L., Smailović, J., and Mozetič, I. (2017). A comparative study of performance estimation methods for time series forecasting. pages 529–538.

Chang, Y.-M., Hsu, N.-J., and Huang, H.-C. (2010). Semiparametric estimation and selection for nonstationary spatial covariance functions. *Journal of Computational and Graphical Statistics*, 19(1):117–139.

Chernozhukov, V., Karl Härdle, W., Huang, C., and Wang, W. (2021). Lasso-driven inference in time and space. *The Annals of Statistics*, 49(3):1702–1735, 34.

Chu, T., Zhu, J., and Wang, H. (2011). Penalized maximum likelihood estimation and variable selection in geostatistics. *The Annals of Statistics*, 39(5):2607–2625, 19.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226.

Cressie, N., Sainsbury-Dale, M., and Zammit-Mangion, A. (2022). Basis-function models in spatial statistics. *Annual Review of Statistics and Its Application*, 9:373–400.

Cressie, N., Shi, T., and Kang, E. L. (2010). Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics*, 19(3):724–745.

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

de Bruin, S., Brus, D. J., Heuvelink, G. B. M., van Ebbenhorst Tengbergen, T., and Wadoux, A. M. J. C. (2022). Dealing with clustered samples for assessing map accuracy by cross-validation. *Ecological Informatics*, 69:101665.

Dey, D., Datta, A., and Banerjee, S. (2021). Graphical Gaussian process models for highly multivariate spatial data. *Biometrika*.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: A bayesian perspective. *Statistica Sinica*, 14(3):731–761.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 34.

Fassò, A., Maranzano, P., and Otto, P. (2022). Spatiotemporal variable selection and air quality impact assessment of covid-19 lockdown. *Spatial Statistics*, 49:100549.

Furrer, R., Bachoc, F., and Du, J. (2016). Asymptotic properties of multivariate tapering for estimation and prediction. *Journal of Multivariate Analysis*, 149:177–191.

Furrer, R., Genton, M. G., and Nychka, D. (2006a). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.

Furrer, R., Genton, M. G., and Nychka, D. (2006b). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.

Gasch, C. K., Hengl, T., Gräler, B., Meyer, H., Magney, T. S., and Brown, D. J. (2015). Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3d + t: The cook agronomy farm data set. *Spatial Statistics*, 14:70–90.

George, E. I., Sun, D., and Ni, S. (2008). Bayesian stochastic search for var model restrictions. *Journal of Econometrics*, 142(1):553–580.

Gibbons, S. and Overman, H. G. (2012). Mostly pointless spatial econometrics? *Journal of Regional Science*, 52(2):172–191.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, 97(458):590–600.

Gneiting, T., Kleiber, W., and Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105(491):1167–1177.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Gonella, R., Bourel, M., and Bel, L. (2022). Facing spatial massive data in science and society: Variable selection for spatial models. *Spatial Statistics*, page 100627.

Hastie, T. (2020). Ridge regularization: An essential concept in data science. *Technometrics*, 62(4):426–433.

Haworth, J. and Cheng, T. (2014). Graphical lasso for local spatio-temporal neighbourhood selection. In *Proceedings the GIS Research UK 22nd Annual Conference. Presented at the GISRUK*, pages 425–433.

Hjorth, U. and Hjort, U. (1982). Model selection and forward validation. *Scandinavian Journal of Statistics*, 9(2):95–105.

Hofierka, J., Parajka, J., Mitasova, H., and Mitas, L. (2002). Multivariate interpolation of precipitation using regularized spline with tension. *Transactions in GIS*, 6(2):135–150.

Hsiang, T. C. (1975). A bayesian view on ridge regression. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 24(4):267–268.

Hsu, N.-J., Chang, Y.-M., and Huang, H.-C. (2012). A group lasso approach for non-stationary spatial–temporal covariance estimation. *Environmetrics*, 23(1):12–23.

Huang, H.-C., Hsu, N.-J., Theobald, D. M., and Breidt, F. J. (2010). Spatial Lasso with applications to GIS model selection. *Journal of Computational and Graphical Statistics*, 19(4):963–983.

Ingram, B. F. and Whiteman, C. H. (1994). Supplanting the 'minnesota' prior: Forecasting macroeconomic time series using real business cycle model priors. *Journal of Monetary Economics*, 34(3):497–510.

Jiang, G. and Wang, W. (2017). Markov cross-validation for time series model evaluations. *Information Sciences*, 375:219–233.

Kang, M. and Katzfuss, M. (2021). Correlation-based sparse inverse Cholesky factorization for fast Gaussian-process inference. *arXiv preprint arXiv:2112.14591*.

Katzfuss, M. and Cressie, N. (2012). Bayesian hierarchical spatio-temporal smoothing for very large datasets. *Environmetrics*, 23(1):94–107.

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555.

Kostov, P. (2010). Model boosting for spatial weighting matrix selection in spatial lag models. *Environment and Planning B: Planning and Design*, 37(3):533–549.

Kostov, P. (2013). Spatial weighting matrix selection in spatial lag econometric model. *Econometrics*, 1(1):20–30.

Krock, M., Kleiber, W., and Becker, S. (2021a). Nonstationary modeling with sparsity for spatial data via the basis graphical lasso. *Journal of Computational and Graphical Statistics*, 30(2):375–389.

Krock, M., Kleiber, W., Hammerling, D., and Becker, S. (2021b). Modeling massive highly-multivariate nonstationary spatial data with the basis graphical lasso. *arXiv preprint arXiv:2101.02404*.

Lam, C. and Souza, P. C. (2016). Detection and estimation of block structure in spatial weight matrix. *Econometric Reviews*, 35(8-10):1347–1376.

Lam, C. and Souza, P. C. (2020). Estimation and selection of spatial weight matrix in a spatial lag model. *Journal of Business & Economic Statistics*, 38(3):693–710.

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., and Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in r. *Journal of Open Source Software*, 4(44):1903.

Lawson, A. B., Carroll, R., Faes, C., Kirby, R. S., Aregay, M., and Watjou, K. (2017). Spatiotemporal multivariate mixture models for bayesian model selection in disease mapping. *Environmetrics*, 28(8):e2465.

Lee, K.-J., Jones, G. L., Caffo, B. S., and Bassett, S. S. (2014). Spatial bayesian variable selection models on functional magnetic resonance imaging time-series data. *Bayesian Analysis (Online)*, 9(3):699.

Lezama Valdes, L.-M., Katurji, M., and Meyer, H. (2021). A machine learning based downscaling approach to produce high spatio-temporal resolution land surface temperature of the antarctic dry valleys from modis data. *Remote Sensing*, 13(22):4673.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 44(1/2):187–192.

Liu, T. (2017). *Model selection and adaptive lasso estimation of spatial models*. Thesis.

Liu, X., Chen, J., and Cheng, S. (2018). A penalized quasi-maximum likelihood method for variable selection in the spatial autoregressive model. *Spatial statistics*, 25:86–104.

Liu, Y. (2022). Adaptive lasso variable selection method for semiparametric spatial autoregressive panel data model with random effects. *Communications in Statistics - Theory and Methods*, pages 1–19.

Ludwig, A., Meyer, H., and Nauss, T. (2016). Automatic classification of google earth images for a larger scale monitoring of bush encroachment in south africa. *International Journal of Applied Earth Observation and Geoinformation*, 50:89–94.

Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies*, 60(3):531–542.

Maranzano, P., Otto, P., and Fassò, A. (2023). Adaptive lasso estimation for functional hidden dynamic geostatistical model. *Stochastic Environmental Research and Risk Assessment*, Accepted for publication on May 1st 2023.

Merk, M. S. and Otto, P. (2021). Directional spatial autoregressive dependence in the conditional first-and second-order moments. *Spatial Statistics*, 41:100490.

Merk, M. S. and Otto, P. (2022). Estimation of the spatial weighting matrix for regular lattice data—an adaptive lasso approach with cross-sectional resampling. *Environmetrics*, 33(1):e2705.

Meyer, H., Milà, C., and Ludwig, M. (2022). *CAST: 'caret' Applications for Spatial-Temporal Models*.

Meyer, H. and Pebesma, E. (2021). Predicting into unknown space? estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9):1620–1633.

Meyer, H. and Pebesma, E. (2022). Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, 13(1):2208.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., and Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software*, 101:1–9.

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. *Ecological Modelling*, 411:108815.

Milà, C., Mateu, J., Pebesma, E., and Meyer, H. (2022). Nearest neighbour distance matching leave-one-out cross-validation for map validation. *Methods in Ecology and Evolution*, n/a(n/a).

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Nandy, S., Lim, C. Y., and Maiti, T. (2017). Additive model building for spatial regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):779–800.

Nychka, D., Wikle, C., and Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2(4):315–331.

Otto, P. and Steinert, R. (2022). Estimation of the spatial weighting matrix for spatiotemporal data under the presence of structural breaks. *Journal of Computational and Graphical Statistics*, (just-accepted):1–44.

Pannu, J. and Billor, N. (2017). Robust group-lasso for functional regression model. *Communications in Statistics - Simulation and Computation*, 46(5):3356–3374.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Pejović, M., Nikolić, M., Heuvelink, G. B., Hengl, T., Kilibarda, M., and Bajat, B. (2018). Sparse regression interaction models for spatial prediction of soil properties in 3d. *Computers & Geosciences*, 118:1–13.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., and Pélissier, R. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11(1):4540.

Pohjankukka, J., Pahikkala, T., Nevalainen, P., and Heikkonen, J. (2017). Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science*, 31(10):2001–2019.

Polson, N. G. and Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 16.

Porcu, E., Bevilacqua, M., and Genton, M. G. (2016). Spatio-temporal covariance and cross-covariance functions of the great circle distance on a sphere. *Journal of the American Statistical Association*, 111(514):888–898.

Porcu, E., Furrer, R., and Nychka, D. (2021). 30 years of space–time covariance functions. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(2):e1512.

R Core Team, . (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, 99(1):39–61.

Reyes, P. E., Zhu, J., and Aukema, B. H. (2012a). Selection of spatial-temporal lattice models: Assessing the impact of climate conditions on a mountain pine beetle outbreak. *Journal of Agricultural, Biological, and Environmental Statistics*, 17(3):508–525.

Reyes, P. E., Zhu, J., and Aukema, B. H. (2012b). Selection of spatial-temporal lattice models: assessing the impact of climate conditions on a mountain pine beetle outbreak. *Journal of agricultural, biological, and environmental statistics*, 17:508–525.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., and Thuiller, W. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, 4:395–421.

Safikhani, A., Kamga, C., Mudigonda, S., Faghih, S. S., and Moghimi, B. (2020). Spatio-temporal modeling of yellow taxi demands in new york city using generalized star models. *International Journal of Forecasting*, 36(3):1138–1148.

Schratz, P., Becker, M., Lang, M., and Brenning, A. (2021). Mlr3spatiotempcv: Spatiotemporal resampling methods for machine learning in r. *arXiv preprint arXiv:2110.12674*.

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. *Ecological Modelling*, 406:109–120.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Schäfer, F., Katzfuss, M., and Owhadi, H. (2021). Sparse Cholesky factorization by Kullback–Leibler minimization. *SIAM Journal on scientific computing*, 43(3):A2019–A2046.

Silverman, B. and Ramsay, J. (2002). *Applied functional data analysis: methods and case studies*. Springer New York, NY.

Simon, N. and Tibshirani, R. (2012). Standardization and the group lasso penalty. *Statistica Sinica*, 22(3):983.

Stein, M. L. (2005). Space–time covariance functions. *Journal of the American Statistical Association*, 100(469):310–321.

Stein, M. L. (2013). Statistical properties of covariance tapers. *Journal of Computational and Graphical Statistics*, 22(4):866–885.

Stein, M. L. (2014). Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statistics*, 8:1–19.

Stein, M. L., Chi, Z., and Welty, L. J. (2004a). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(2):275–296.

Stein, M. L., Chi, Z., and Welty, L. J. (2004b). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):275–296.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4):437–450.

Telford, R. J. and Birks, H. J. B. (2005). The secret assumption of transfer functions: problems with spatial autocorrelation in evaluating model performance. *Quaternary Science Reviews*, 24(20):2173–2179.

Telford, R. J. and Birks, H. J. B. (2009). Evaluation of transfer functions in spatially structured environments. *Quaternary Science Reviews*, 28(13):1309–1316.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Trachsel, M. and Telford, R. J. (2016). Technical note: Estimating unbiased transfer-function performances in spatially structured environments. *Clim. Past*, 12(5):1215–1223.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2019). blockcv: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2):225–232.

van de Wiel, M. A., Te Beest, D. E., and Münch, M. M. (2019). Learning from a lot: Empirical bayes for high-dimensional model-based prediction. *Scandinavian Journal of Statistics*, 46(1):2–25.

van Erp, S., Oberski, D. L., and Mulder, J. (2019). Shrinkage priors for bayesian penalized regression. *Journal of Mathematical Psychology*, 89:31–50.

van Wieringen, W. N. (2015). Lecture notes on ridge regression. *arXiv preprint arXiv:1509.09169*.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):297–312.

Vecchia, A. V. (2018). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):297–312.

Ver Hoef, J. M., Hanks, E. M., and Hooten, M. B. (2018). On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models. *Spatial statistics*, 25:68–85.

Wang, H. and Zhu, J. (2009). Variable selection in spatial regression via penalized least squares. *Canadian Journal of Statistics*, 37(4):607–624.

Wang, Y., Finazzi, F., and Fassò, A. (2021). D-stem v2: A software for modeling functional spatio-temporal data. *Journal of Statistical Software*, 99(10):1 – 29.

Wen, Y., Shen, X., and Lu, Q. (2018). Genetic risk prediction using a spatial autoregressive model with adaptive lasso. *Statistics in Medicine*, 37(26):3764–3775.

Wikle, C. K., Berliner, L. M., and Cressie, N. (1998). Hierarchical bayesian space-time models. *Environmental and Ecological Statistics*, 5(2):117–154.

Wolpert, D. H. and Strauss, C. E. M. (1996). *What Bayes has to Say about the Evidence Procedure*, pages 61–78. Springer Netherlands, Dordrecht.

Wood, S. N. (2017). *Generalized additive models: an introduction with R.* chapman and hall/CRC, 2nd edition.

Xiao, Y., Gu, X., Yin, S., Shao, J., Cui, Y., Zhang, Q., and Niu, Y. (2016). Geostatistical interpolation model selection based on arcgis and spatio-temporal variability analysis of groundwater level in piedmont plains, northwest china. *SpringerPlus*, 5(1):425.

Ye, J., Lazar, N. A., and Li, Y. (2011). Sparse geostatistical analysis in clustering fmri time series. *Journal of neuroscience methods*, 199(2):336–345.

Zapata, J., Oh, S.-Y., and Petersen, A. (2022). Partial separability and functional graphical models for multivariate Gaussian processes. *Biometrika*, 109(3):665–681.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*.

Zeng, X. and Martinez, T. R. (2000). Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(1):1–12.

Zhang, H., Huang, X., Gan, J., Karmaus, W., and Sabo-Attwood, T. (2016). A two-component g-prior for variable selection. *Bayesian Analysis*, 11(2):353–380, 28.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zhao, Y. and Karypis, G. (2002). Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 515–524.

Zhu, J., Huang, H.-C., and Reyes, P. E. (2010). On selection of spatial linear models for lattice data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):389–402.

Zhu, Z. and Liu, Y. (2009). Estimating spatial covariance using penalised likelihood with weighted l_1 penalty. *Journal of Nonparametric Statistics*, 21(7):925–942.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509–1533.