Statistics Surveys
Vol. 18 (2024) 299–340
ISSN: 1935-7516
https://doi.org/10.1214/24-SS150

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

Philipp Otto[1], **Alessandro Fassò**[2] **and Paolo Maranzano**[3,4]

[1] *School of Mathematics and Statistics, University of Glasgow, United Kingdom, e-mail:* philipp.otto@glasgow.ac.uk, *corresponding author*

[2] *University of Bergamo, Italy, e-mail:* alessandro.fasso@unibg.it

[3] *Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca, Italy, e-mail:* paolo.maranzano@unimib.it

[4] *Fondazione Eni Enrico Mattei (FEEM), Italy*

Abstract: This review article focuses on regularised estimation procedures applicable to geostatistical and spatial econometric models and suitable cross-validation techniques for spatiotemporal data. These methods are particularly relevant in the case of big geospatial data for dimension 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. Guided by the elements of a mixed-effects model setup, which allows for a variety of spatiotemporal models, we show different regularisation procedures and how they can be used for the analysis of geo-referenced data, e.g. for selection of relevant regressors, dimension reduction of the covariance matrices, detection of conditionally independent locations, or the estimation of a full spatial interaction matrix. The second part is dedicated to cross-validation strategies, which are important for evaluating the model performance and selecting regularisation parameters. We outline the three key assumptions a cross-validation partitioning needs to fulfil and discuss how this can be achieved for time series, spatial data and spatiotemporal data. Additionally, software implementations for these techniques are discussed.

MSC2020 subject classifications: Primary 62-02, 62H11, 62J07; secondary 62M45, 62P12, 62K99.
Keywords and phrases: Big data, dimension reduction, geostatistics, spatial econometrics.

Received May 2024.

1. Introduction

In the era of big geospatial data, analysing intricate spatial and spatiotemporal processes has become increasingly vital and challenging. This is attributed to the

arXiv: 2402.00183

growing volume and finer resolution of geo-referenced data, including remotely sensed, crowd-sourced or 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 dimension 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.

This review paper focuses on regularised estimation procedures for spatiotemporal continuous data, mainly focusing on Gaussian geostatistical models and spatial autoregression models, which are mostly applied for data on regular or irregular spatial grids, e.g., areal data. The latter ones are often called *spatial econometrics models*, but we will refer to them as *spatial autoregression models* because (a) the response variable is typically explicitly correlated with its adjacent observations in an autoregressive manner, and (b) they are not only tied to applications in economics, neither do they originate from economics/econometrics. These models are typically used for data defined on a discrete set of spatial locations, e.g., regular or irregular lattices, including polygons such as municipalities, counties or countries, making them attractive modelling approaches in econometrics (see, e.g., Billé and Arbia 2019; Anselin 2010; Arbia 2016), ecology (see, e.g., Ver Hoef et al. 2018; Lichstein et al. 2002), or epidemiology (see, e.g., Lee 2011; Gebreab, Duncan and Kawachi 2018). Instead, in geostatistics, spatial dependence is modelled using covariance functions depending on the distance between two observations[1]. This makes the model attractive for processes on a continuous space and data which is irregularly observed across space, such as air pollution at ground-level measurement stations. Spatial point processes, such as Poisson processes, will not be the focus of this review paper (see González et al. 2016 for a review on spatial point processes).

Regularised estimation procedures offer effective solutions for these challenges from multiple perspectives. For instance, they can be used for model selection and dimension 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 autoregression/econometrics models (Section 4). Since the degree of regularisation is typically chosen based on pre-

[1]Note that the distance is not necessarily measured as Euclidean distance, but different distance measures such as great-circle distances for modelling processes on a sphere or suitable network distances for processes on networks can be considered.

dictive accuracy, we also review cross-validation techniques which are suitable under spatiotemporal dependence (Section 5). Finally, Section 7 concludes the article and discusses future research directions.

2. General framework

Assume that we are observing a continuous q-dimensional vector over space and time. The related q-variate dataset may be formalised as

$$\{y_{t_j}(\boldsymbol{s}_i) \in \mathbb{R}^q : i = 1, \ldots, N, j = 1, \ldots, T\} \tag{1}$$

where $\boldsymbol{s}_i$ are points in the spatial domain $D_{\boldsymbol{s}}$, which may be a Euclidean or non-Euclidean space. Common choices are the plain $\mathbb{R}^2$, the Earth sphere $\mathbb{S}^2$ or a discrete grid of points. The time domain, say $D_t \ni t_j$, is assumed discrete and, ignoring missing values, is made by equidistant time points. For simplicity, we use the set of integers, $D_t = \mathbb{Z}$ and $t_j = j = 1, \ldots, T$. In the statistical framework, data $y_t(\boldsymbol{s}_i)$ are assumed to be generated by the spatiotemporal stochastic process $\{Y_t(\boldsymbol{s}) \in \mathbb{R}^q : \boldsymbol{s} \in D_{\boldsymbol{s}}, t \in D_t\}$, (see Cressie and Wikle, 2015). A continuous temporal domain is sometimes considered in geostatistics, see, e.g., Porcu and White (2022), but we focus on the first case, which may be called the spatial time series approach.

The above-mentioned spatial and temporal domains can assume several forms, resulting in multiple data categories. In particular, following (Wikle, Zammit-Mangion and Cressie, 2019, Section 2.1), we can distinguish among three main classes of spatiotemporal data. By geostatistical data, we mean phenomena which can be measured at continuous locations over a given spatial domain. Typical examples are air quality, weather and climate measurements, such as the PM_{10} concentrations reported in Figure 1. By areal or lattice data, we mean a phenomenon defined on a finite or countable subset in space over a specific time span. Examples of areal data are easily found in the socio-economic, political or medical-epidemiological realms, where observations are often reported by area (e.g., municipalities or regions). For example, in Figure 2, we show the evolution of per capita income for the European provinces (NUTS-3 classification) between 2011 and 2020. They provide a common playground for statisticians, econometricians, or applied social scientists, as the neighbourhood structure among units can provide relevant insights in many empirical contexts while preserving the interpretability of the phenomenon.

There are different sources of correlation 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 earlier but less similar to the temperature two days ago.
   Also, more complex time-related structures are possible. For example, considering periodicity, the daily and seasonal cycles often require non-monotonic correlation structures.

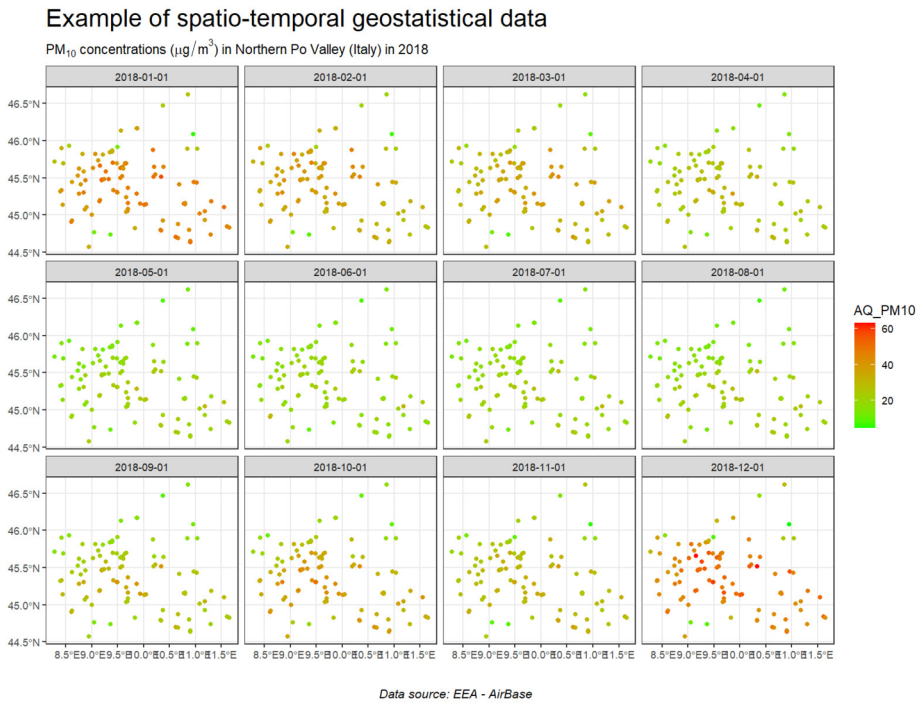

FIG 1. *Illustrative example of real-world geostatistical spatiotemporal data. The plot shows the observed PM_{10} concentrations in 2018 at the 101 locations in Northern Italy considered by Fassò et al. (2023). For each panel, the observed average monthly concentrations are reported.*

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. For example, 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. Univariate models may be sufficient when the q variables are truly independent, and applying a multivariate model, in this case, could introduce spurious correlations and unnecessary complexity. On the other hand, if the variables exhibit shared spatial or temporal structure or latent dependencies, a multivariate approach can capture these relationships and provide more efficient joint inference. Careful consideration of the underlying data structure is necessary to determine the most appropriate modelling strategy.
   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, Datta and Banerjee, 2022; Zapata, Oh and Petersen, 2022, for graphical Gaussian processes for multivari-

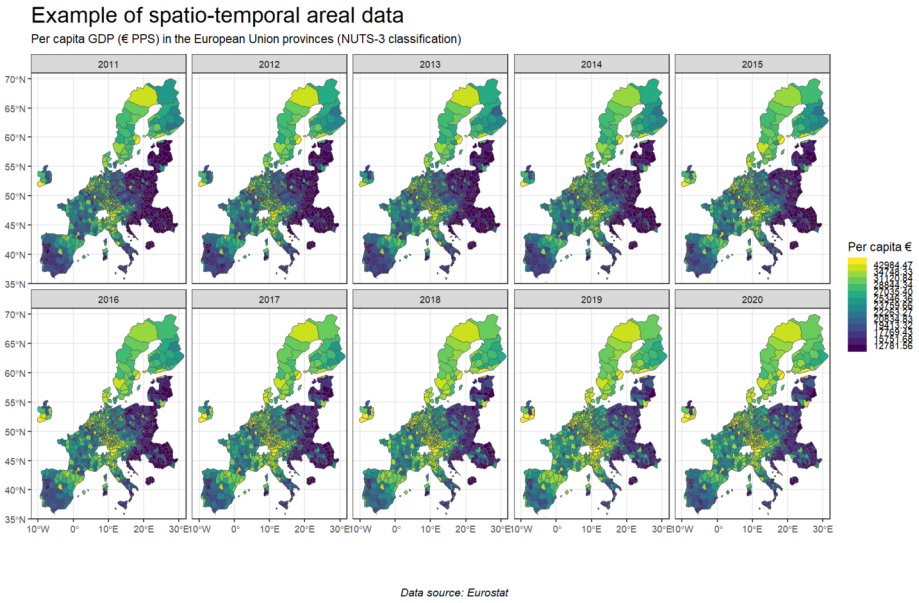

FIG 2. *Illustrative example of spatiotemporal areal (irregular lattice) data. The plot shows the estimated yearly per capita gross domestic product (€ per capita) in the 1088 European Union provinces (NUTS-3 classification) from 2011 to 2020. Data source: Eurostat.*

ate spatial data). However, graphical models are beyond the scope of this paper.

There are two alternative, complementary 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 nearby observations. 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 and regularised estimation

A quite general frame for modelling data in (1) is given by the mixed-effects spatiotemporal model, which is given by

$$Y_t(\boldsymbol{s}) = \mu_t(\boldsymbol{s}) + \omega_t(\boldsymbol{s}) + \varepsilon_t(\boldsymbol{s}), \tag{2}$$

where $\mu_t(\boldsymbol{s})$ is the fixed-effects component, $\omega_t(\boldsymbol{s})$ is the random-effects component, and $\varepsilon_t(\boldsymbol{s})$ is the stochastic error at time t and location $\boldsymbol{s}$, which is typically assumed to follow a (Gaussian) white noise process. The fixed-effects

term typically models the mean behaviour and the influence of exogenous variables, potentially implying a non-stationary behaviour in the mean of $Y_t(\boldsymbol{s})$. Furthermore, the random-effects term is a weakly stationary random process, accounting for any additional dependence in the data, including temporal and spatial dependence, heterogeneity, and cross-correlation between the variables. We will follow this distinction throughout the paper and discuss regularised estimation procedures for the parameters of $\mu_t(\boldsymbol{s})$ and $\omega_t(\boldsymbol{s})$. Remembering that a weakly stationary process has a constant mean across all locations and time points, and the covariance is only a function of the difference of time indices and spatial location coordinates, the process $Y_t(\boldsymbol{s})$ is trend stationary.

Penalised estimation procedures can be used to shrink certain model parameters towards a pre-specified target. Generally speaking, the idea of these estimators is to balance the in-sample model fit and the distance to the user-defined target, called "penalty". The former usually improves with increasing model flexibility, whilst the latter restricts the model flexibility, usually leading to an inferior in-sample model fit. In other words, the penalty term serves to (automatically) adjust the model structure towards a certain user-defined target. The general form of a regularised estimation can be expressed as

$$\hat{\theta} = \arg\min_{\theta} \left[\mathcal{L}(\boldsymbol{\theta}; \mathbf{X}, \mathbf{y}) + P(\boldsymbol{\theta}, \boldsymbol{\theta}_0, \boldsymbol{\lambda}) \right], \tag{3}$$

where $\mathcal{L}(\boldsymbol{\theta}; \mathbf{X}, \mathbf{y})$ represents the model fit term, e.g. the error sum of squares or the negative log-likelihood function. Moreover, $\mathbf{X}$ is the spatiotemporal set of potential covariates, $\mathbf{X}_t(\boldsymbol{s_i}), i = 1, \ldots, N, t = 1, \ldots, T$ entering in the fixed-effects component $\mu_t(\boldsymbol{s})$, and $\mathbf{y}$ is the set of all observations of $Y_t(\boldsymbol{s_i})$ in Equation 2. The penalty term $P(\boldsymbol{\theta}, \boldsymbol{\theta}_0, \boldsymbol{\lambda})$ introduces regularisation and includes a tuning parameter $\boldsymbol{\lambda}$, which could be scalar but also a vector, and the so-called shrinkage target $\boldsymbol{\theta}_0$. The form of $P(\boldsymbol{\theta}, \boldsymbol{\theta}_0, \boldsymbol{\lambda})$ can vary, allowing for a range of regularisation schemes. For instance, in the case of convex penalties, the LASSO (least absolute shrinkage and selection operator) imposes an ℓ_1-norm penalty of the differences $\boldsymbol{\theta} - \boldsymbol{\theta}_0$. Choosing the shrinkage target $\boldsymbol{\theta}_0$ equal to a vector of zeros encourages sparsity in the estimated parameters. In the case of graphical models, such as the graphical LASSO, the shrinkage target may be a diagonal matrix $\mathbf{I}$, encouraging the estimated covariance matrix to have fewer off-diagonal non-zero elements, reflecting conditional independence relationships. Non-convex penalties, such as the smoothly clipped absolute deviation (SCAD, Fan and Li 2001) or the minimax concave penalty (MCP, Zhang 2010), are also frequently employed to allow for more flexible shrinkage behaviour, with less bias introduced for large parameter estimates.

As mentioned above, probably, the most often applied shrinkage target is zero. In this case, the parameters are shrunk towards zero, which means that a parameter is excluded from the model if its estimate is equal to the zero target. Thus, these methods are suitable for model selection (simultaneous parameter estimation and model selection). In particular, when applied to the fixed-effects part, they can automatically select relevant regressors. Compared to standard model selection procedures, such as step-wise selection based on cross-validation

goodness-of-fit measures, regularised estimation procedures are typically computationally more efficient[2].

For the rest of the paper, we will mainly focus on Gaussian models. If the data is continuous but non-Gaussian, a common approach in spatiotemporal statistics is to apply suitable (non-linear) transformations of the observed process, e.g., logarithmic, Box-Cox, or square-root transformations, such that the transformed process follows a Gaussian distribution (see also Wallin and Bolin, 2015). Moreover, a transport map $\mathcal{T}(Y_t(\boldsymbol{s}))$ can be used to transform a continuous multivariate distribution into a specific target distribution, e.g., a Gaussian distribution. Katzfuss and Schäfer (2024) proposed a scalable Bayesian transport map for spatiotemporal data to handle non-Gaussian data, which also enables regularisation of the mapping by an appropriate choice of the prior distributions to obtain sparse transport maps (see also Wiemann and Katzfuss, 2023). Spatiotemporal models can also be explicitly designed for categorical or count data, e.g., logistic or Poisson spatial models for epidemiological data (Paciorek, 2007) often in the fashion of generalised linear or additive models (cf. Wood and Augustin, 2002; Aswi et al., 2019; Smart et al., 2010; Chattopadhyay and Deb, 2024; Otto et al., 2024), giving penalised likelihood estimation procedures in the frame of Equation 3 (but outside our mainly considered setting in (2)).

Generally, there are two main sources of increasing computational complexity: model dimension and data dimension. The former is related to the fixed-effects component when considering variable selection. This is developed in Section 3.1. Moreover, the model dimension is related to the random-effects component when modelling spatial and temporal correlations, which are considered in Section 3.2.

Considering data dimension, the observation covariance matrix of the data set (1) is $qNT \times qNT$ dimensional, and its brute-force inversion has a cubic computational cost. Hence, specialised modelling and computing techniques are needed. The leading term in modern large spatial data sets is the number of spatial locations N. Fortunately, the correlation often decreases with an increasing distance between observations, allowing for a sparse covariance or precision matrix representation. Regularised estimation procedures with a shrinkage target of zero can be used to introduce zeros in the spatial covariance matrix or the precision matrix, indicating conditional independence between these observations. This will be the subject of Section 3.2. Similarly, if the number of variables q increases, we observe a corresponding cubic increase characterised by the fact that the cross-correlation matrices for contemporaneous and colocated observations are usually dense. On the other side, the temporal dimension is often less critical because causal time series models may leverage on the fact that only past observations influence future observations, and the computational complexity may be reduced to be linear in T, see e.g. Wang, Finazzi and Fassò (2021), §2.4. These issues are also considered in Section 3.2.

[2]Note that only one or a few hyper-parameters, i.e., regularisation parameters, need to be selected via cross-validation, while for other complete or stepwise selection procedures, significantly more cross-validation iterations are required.

3.1. Regularisation for the fixed-effects component

Linear regression models, whether referring to spatiotemporal or unstructured data, can be estimated using various techniques, including Likelihood maximisation and least squares. However, in high-dimensional settings, traditional methods for regression models might not be directly applied (Nandy, Lim and Maiti, 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. Specifically, we refer to the task of selecting the most relevant predictors among a large set of candidates.

The state-of-the-art statistical literature inherent to variable selection 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, Ritov and Tsybakov, 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, Chi and Welty, 2004; 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, Zhu and Wang (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, Genton and Nychka, 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 the classic LASSO algorithm does not provide selection-consistent estimates when predictors are correlated. 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).

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) (including also non-convex penalties like SCAD), spatial autoregressive models (Cai and Maiti, 2020), spatial error models (Al-Momani, Hussein and Ahmed, 2017), regression models with spatially dependent data (Huang et al., 2010) and conditional autoregressive models (Gonella, Bourel and Bel, 2022). Also, when considering additive spatial models with potential non-linear effects, weighted versions of penalised least squares can be applied (Nandy, Lim and Maiti, 2017). Cai et al. (2019) proposed a generalised method-of-moments (GMM) 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, 1988). Eventually, several application-oriented papers combine classic LASSO approaches and geostatistical models in multi-step procedures (e.g., Fassò, Maranzano and Otto, 2022; Ye, Lazar and Li, 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, Hsu and Huang, 2010, for group-LASSO approaches in this context). For instance, Maranzano, Otto and Fassò (2023) use a PMLE with an adaptive LASSO penalty to select relevant functional covariates or their statistically relevant regions using the hidden dynamic geostatistical model. Eventually, Hsu, Chang and Huang (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 the random-effects component

Suppose that the random-effects component follows a stationary q-variate Gaussian process, i.e.,

$$\{\omega_t(\boldsymbol{s}) : \boldsymbol{s} \in D_{\boldsymbol{s}}, t \in D_t\} \quad \sim \quad N_p(0, C_\theta(\boldsymbol{s} - \boldsymbol{s}', t - t')), \tag{4}$$

where C_θ is a matrix covariance function that depends on the difference between any two locations $\boldsymbol{s}$ and $\boldsymbol{s}'$ and two arbitrary time points t and t'. It is the fundamental building block of the covariance matrix of the data set (1) and, for $q > 1$, also provides the cross-covariances between the components of the observation vector y. In order to have a positive semidefinite observation covariance matrix, C_θ must be a *valid* covariance function. See Gneiting 2002; Gneiting, Kleiber and Schlather 2010; Stein 2005; Nychka, Wikle and Royle 2002; Porcu, Bevilacqua and Genton 2016 or Porcu, Furrer and Nychka 2021 for a historical review of space-time covariance functions. The covariance function in (4) usually depends on some unknown parameters θ, which have to be estimated.

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, Genton and Nychka, 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, and the use of tapering in the multivariate case is considered in Bevilacqua et al. (2016). This method can also be applied in likelihood-based estimation procedures (see Kaufman, Schervish and Nychka 2008, and Furrer, Bachoc and Du 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, Kleiber and Becker (2021) introduced a graphical LASSO procedure to induce zeros in the precision matrix of a spatial process. Krock et al. (2023) extended this approach to be used in the multivariate case ($q > 1$).

Moreover, since the covariance matrix is a positive definite matrix by definition, 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, Chi and Welty (2004) 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, Katzfuss

and Owhadi (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, Shi and Kang (2010) (spatiotemporal fixed-rank filtering). Back to the subject of this review paper, Chang, Hsu and Huang (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, Sainsbury-Dale and Zammit-Mangion 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, Chang and Huang (2012) suggested penalised procedures to choose these local basis functions. Furthermore, Kang and Katzfuss (2023) 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, Oberski and Mulder (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 *fully Bayesian* or *hierarchical Bayesian* (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, Oberski and Mulder, 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* approach (van de Wiel, Te Beest and Münch, 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, Oberski and Mulder (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 spatiotemporal 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, Berliner and Cressie, 1998) and an empirical Bayes approach (Fahrmeir, Kneib and Lang, 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 model where

dimension 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, Sun and Ni, 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, York and Mollié, 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 spatiotemporal disease mapping (Carroll et al., 2016b). Among others, relevant advantages include getting an automatic final model fit.

4. Spatial autoregression

For spatial autoregression, the outcome variable is explicitly correlated with adjacent observations, which also induces a certain structure in the covariance matrix. Thereby, the local neighbourhood is defined by a suitable $N \times N$ weight matrix $\mathbf{W}$, which is weighting all observations, such that the product of $\mathbf{W}$ and the dependent variable at time t is the weighted average of the contemporaneous adjacent observations. For instance, a common choice for areal data (e.g., economic county-level data such as regional GDPs or housing prices) is to assume the spatial weight matrix as a row-standardised contiguity matrix. The (i, j)-th element of a spatial contiguity is equal to one if location i and j share a common border and zero otherwise. After the row standardisation (i.e., all elements are divided by the corresponding row sum), the product of $\mathbf{W}$ and a variable of interest is equal to the sample average of all adjacent regions, facilitating its interpretation. Generally speaking, the weight matrix can also be viewed as an adjacency matrix in network modelling, giving the relation between two different nodes/locations/entities.

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 we can consider a dynamic spatiotemporal autoregressive panel model (Yu, De Jong and Lee, 2008) as a starting point in this Section. This model is given by

$$\boldsymbol{Y}_t = \mathbf{X}_t\boldsymbol{\beta} + \rho\mathbf{W}\boldsymbol{Y}_t + \delta\boldsymbol{Y}_{t-1} + \gamma\mathbf{W}\boldsymbol{Y}_{t-1} + a_t\mathbf{1}_n + \boldsymbol{c} + \boldsymbol{\varepsilon}_t \qquad \text{for } t = 1, \ldots, T, \tag{5}$$

where $\mathbf{X}_t$ is a matrix of regressors at time t, $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients, a_t are temporal fixed effects (constant across space), $\boldsymbol{c}$ is a vector of spatial fixed effects (constant over time), and ρ, δ, γ are the spatial, temporal and spatiotemporal autoregressive parameters, respectively. Moreover, $\boldsymbol{\varepsilon}_t$ is the vector of white noise model errors.

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

$$\begin{aligned}
\boldsymbol{\mu}_t &= (\mu_t(\boldsymbol{s}_1), \ldots, \mu_t(\boldsymbol{s}_n))' \\
&= (\mathbf{I} - \rho\mathbf{W})^{-1}(\mathbf{X}_t\boldsymbol{\beta} + \delta\boldsymbol{Y}_{t-1} + \gamma\mathbf{W}\boldsymbol{Y}_{t-1} + a_t\mathbf{1}_n + \boldsymbol{c}), \\
\boldsymbol{\omega}_t + \boldsymbol{\varepsilon}_t &\sim N_n(\mathbf{0}, (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\Sigma}_\varepsilon(\mathbf{I} - \rho\mathbf{W}')^{-1})
\end{aligned}$$

with $\boldsymbol{\omega}_t = (\omega_t(\boldsymbol{s}_1), \ldots, \omega_t(\boldsymbol{s}_n))'$ and $\boldsymbol{\varepsilon}_t = (\varepsilon_t(\boldsymbol{s}_1), \ldots, \varepsilon_t(\boldsymbol{s}_n))'$. That is, the (spatially correlated) random effects model and the model errors are jointly modelled as a Gaussian distribution with a specific structure of the covariance matrix implied by the spatial weights. The covariance matrix of the error term is denoted by $\boldsymbol{\Sigma}_\varepsilon$, typically diagonal and or a multiple of the identity matrix in the homoscedastic case. It is worth noting that temporal dependence is implied due to the temporal autoregressive structure in $\boldsymbol{\mu}_t$. Moreover, setting $\delta = \gamma = 0$, we obtain a simple spatial autoregressive model (SAR), which could also be applied in a purely spatial setting. A popular alternative specification of the covariance structure is implied for conditional autoregressive (CAR) models, where

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

For this approach, additional assumptions on the weight matrix are needed to ensure a well-defined covariance matrix, e.g., symmetry and positive eigenvalues of $\mathbf{W}$. Generally, there is a relation between geostatistical and spatial autoregressive models, and both approaches are equivalent under certain conditions (Ver Hoef, Hanks and Hooten, 2018, Theorem 1). Below, our focus will be mainly on Gaussian models, but for the sake of completeness, it is worth noticing that the logic of these models can also be transferred to other non-Gaussian cases, e.g., modelling count data (Congdon, 2022), binary data (Calabrese and Elkink, 2014), or compositions (Thomas-Agnan et al., 2021).

4.1. Regularisation for the mean model

For spatial autoregressive (SAR) and conditional autoregressive (CAR) models, Gonella, Bourel and Bel (2022) proposed a LASSO estimation procedure

to select the relevant covariates in the regression term. Among others, Wen, Shen and Lu (2018) use a spatial autoregressive model with adaptive LASSO penalty to detect relevant autoregressive parameters in genetic studies. Zhu, Huang and Reyes (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, Zhu and Aukema (2012) 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, Hsu and Huang (2010) and Hsu, Chang and Huang (2012) additionally reduced the model's complexity by combining covariance tapering and PMLE for spatial and spatiotemporal settings, respectively.

An alternative spatial LASSO approach has been proposed by Samarov, Hwang and Litorja (2015). They assume a standard linear model with spatially varying coefficients, where spatially adjacent coefficients should have similar estimates. For this reason, they propose an additional spatially weighted penalty term in the LASSO regression. More precisely, the squared distance penalty between two adjacent coefficients $\boldsymbol{\beta}(\boldsymbol{s}_i)$ and $\boldsymbol{\beta}(\boldsymbol{s}_j)$ at locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, weighted by w_{ij}, prevents abruptly varying coefficients of nearby locations. As Samarov, Hwang and Litorja (2015) showed for the analysis of hyperspectral images, these weights can also include spectral information of the image.

4.2. Regularisation for the spatial dependence structure

The precision matrix of such spatial autoregressive models is given by

$$(\mathbf{I} - \rho\mathbf{W})'\boldsymbol{\Sigma}_{\varepsilon}^{-1}(\mathbf{I} - \rho\mathbf{W}) \tag{6}$$

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 specific 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, Zhu and Aukema (2012) 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, Zhu and Aukema (2012) also penalised the regressive parameters using a second penalty term. Similarly, Liu, Chen and Cheng (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 (2023). 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 $\hat{\beta}_\lambda$ are biased but are more efficient, i.e., their variability is smaller. This is known as the bias-variance trade-off (see, for instance, Section 2.9 of Hastie, Tibshirani and Friedman, 2009). Additionally, opting for a λ value that is too small can result in overfitting, while selecting a value that is too large can lead to underfitting (see, e.g., Boonstra, Mukherjee and Taylor, 2015). The optimal penalty parameter is usually selected based on model fit or out-of-sample predictive performance. Other 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 $(\sigma^2_\varepsilon, \lambda)'$. 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, Oberski and Mulder, 2019). Alternatively, Otto and Steinert (2023) 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 may be assessed either in-sample or, better, out-of-sample, typically using cross-validation in terms of the distance between the observed and predicted values. 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 while providing guidelines to choose the suitable cross-validation procedure according to the specificities of the data, including some brief remarks on dependent data and the problem at hand (e.g., model identification or model selection).

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 and test sets.

Regarding the first point, if a model were trained on peak points of seasonal time series and tested on the valley points, the prediction errors would be over-

estimated. Thus, while periodic partitioning should be avoided, a random partitioning strategy should always be preferred. For illustration of the other two points, consider overlapping folds or partitions. In such a situation, training and test sets would be mutually correlated, as well as the resulting training and test errors. As a direct consequence, the sample variance estimator would remarkably underestimate the actual variance of test errors. Independence among the sets can be assured by leaving a certain *distance* between training and test samples. Thus, for a given test set, all other correlated samples have to be removed from the training set to avoid overfitting It means, for example, that when considering spatial or temporal data, the nearby measurements (locations in space or time points) should be removed from the test set when validating a point.

Depending on the data structure (e.g., cross-sectional, clustered, spatial, temporal data), as well as on modelling purposes and data-specific features (e.g., seasonality or non-Gaussianity), the cross-validation schemes to be adopted may vary significantly (Arlot and Celisse, 2010; Hewamalage, Ackermann and Bergmeir, 2023). In the case of independent data, classical CV schemes, such as random k-fold, stratified k-fold (Zeng and Martinez, 2000; Ludwig, Meyer and Nauss, 2016) or Generalized CV (Boonstra, Mukherjee and Taylor, 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, Heath and Wahba, 1979). For spatially and temporally dependent data, standard random k-fold cross-validation procedures should not be applied. As shown by Schratz et al. (2019), while performances are overestimated 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 spatiotemporal 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, Katurji and Meyer, 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).

5.1. Cross-validation in time

Cross-validation is usually applied for stationary time series by segmenting the time series into contiguous subsets to preserve the temporal correlation structure characterising the observations. Two main strategies are commonly used: (1) k-fold blocked subsets CV, where each subset is treated as the test set in turn, and the rest are used for training and forecasting (Bergmeir and Benítez, 2012; Bergmeir, Costantini and Benítez, 2014; Bergmeir, Hyndman and Koo, 2018); and (2) last-block CV or forward validation CV, where only the final block is the test set, and previous blocks are used for training (Hjorth, 1982). The key distinction between the two approaches lies in preserving temporal order. Indeed, while the latter preserves the natural order of temporal observations by forecasting future values using past observations (recall that the model is never tested on past data relative to the training data), the former uses both past and future values to predict the current test set. In both cases, 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 3. Further CV strategies for temporal data can be found in Jiang and Wang (2017) (Markov-CV and partitioned-CV), Cerqueira, Torgo and Mozetič (2020) (prequential approach), and Cerqueira et al. (2017) (Monte Carlo replications last-block approach).

Moreover, time series data are often characterised by non-stationarity of several typologies (Hewamalage, Ackermann and Bergmeir, 2023). Sources of non-stationarity are seasonality, trends (both deterministic and stochastic), structural breaks or heteroskedasticity. 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. Potential meaningful data splitting strategies include the use of a weighted overlapping approach in which the whole series is used in training and testing steps (Bergmeir and Benítez, 2012) or the use of out-of-sample repeated holdout procedures applied in multiple testing periods (Cerqueira, Torgo and Mozetič, 2020). For an extensive discussion of the role of partitioning schemes for temporal data in a forecasting context, see the recent paper by Hewamalage, Ackermann and Bergmeir (2023) in which detailed guidelines on the most appropriate strategy to implement based on the problem to be addressed and the relevant characteristics of the data at hand are provided.

5.2. Cross-validation in space

While spatial data share similar challenges related to autocorrelation with time series, the former involves at least two-dimensional coordinates, whereas time

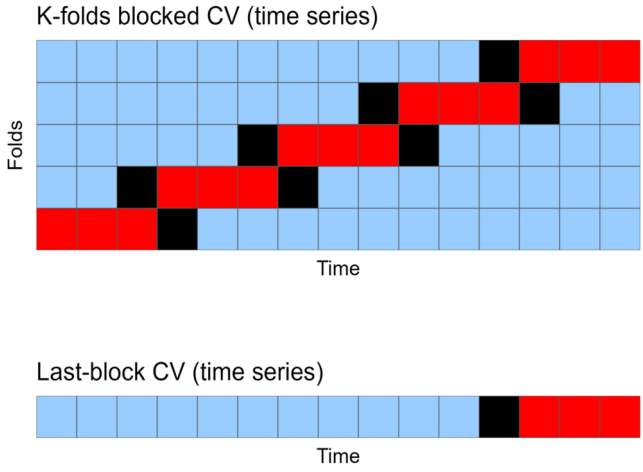

FIG 3. *Temporal CV schemes. For a given iteration fold, blue blocks represent time points in the training set, and red blocks represent time points in the test set. Black blocks are the buffer points omitted from both the training and test sets.*

is one-dimensional (Roberts et al., 2017). 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 also accounts 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).

To address the above issues, spatial CV methods and spatial variable selection techniques, like recursive and forward spatial feature selection, can be employed (Meyer et al., 2018, 2019). 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 between the training and test data (Roberts et al., 2017). In Figure 4, 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 Le Rest et al. (2014), 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 con-

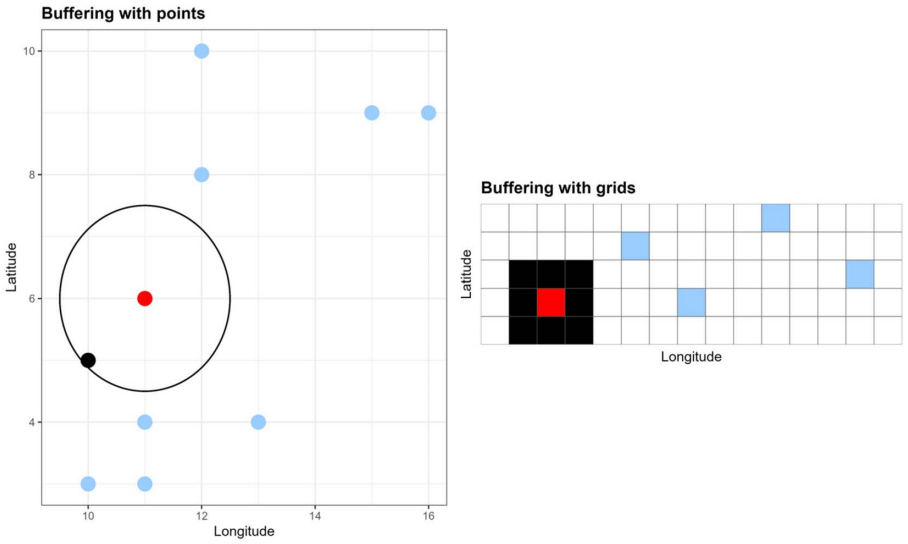

FIG 4. *Spatial buffering for point data (left panel) and for gridded data (right panel). The red points represent the test locations, while the blue points represent training locations. Black points represent locations within the buffer of the test set, which are excluded from both the training and test.*

text. 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 fitting 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 locations 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, Finazzi and Fassò (2021) proposed a heuristically modified k-means algorithm favouring partitions with a similar number of elements. One may also consider the inverse sampling-intensity weighting system proposed 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. Even if not explicitly built for treating spatial and temporal CV tasks, potential extensions of the above-cited clustering approaches include the hierarchical spatiotemporal clustering with spatial constraints (Chavent et al., 2018) and spatially-clustered regression (Sugasawa and Murakami, 2021).

5.3. Cross-validation in both space and time

When spatiotemporal 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 with respect to the spatial blocks and considers the whole time series of each location. LLTO 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 5, we show a schematic example of the three above-mentioned target-oriented CV schemes for spatiotemporal data. Time stamps are reported on the rows for each panel, while spatial locations are on the columns. Regarding the latter, assume that each column represents a pair of longitude and

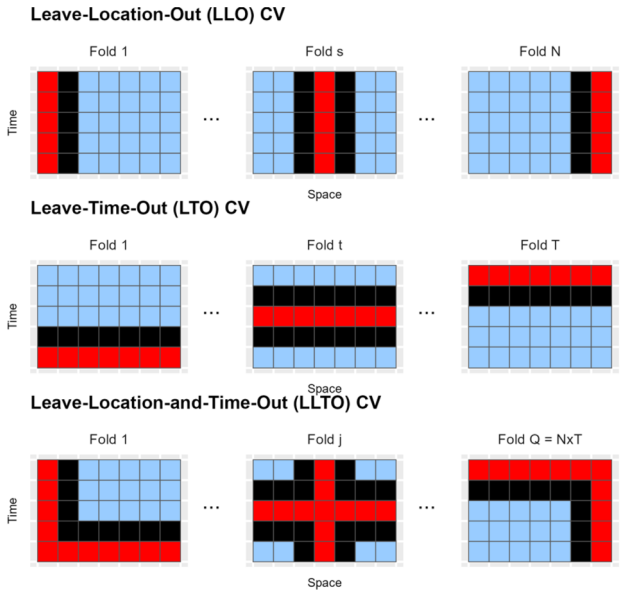

FIG 5. *Schematic matrix representation of the spatiotemporal target-oriented CV schemes. In the first row, the LLO algorithm iterates over the columns; in the second row, the LTO algorithm iterates over rows; in the third row, the LLTO algorithm iterates over spatiotemporal cells. Blue cells represent values used for training the model, while red cells represent points in the test set. Black blocks are observations lying within the spatiotemporal buffer that separate the test set from the training set.*

latitude values and that stations are ordered according to some distance criterion. Cells marked in red represent values used in the test set, blue cells are the values used to train the model, and black cells are the buffering values used to separate training and test sets. According to the chosen scheme, the algorithm iterates over the rows (LTO), over the columns (LLO) or over the cells (LLTO). A geographical representation of the three algorithms is depicted in Figure 6, in which longitude and latitude are on the x-axis and y-axis, respectively, while each time stamp defines a specific panel.

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. However, as the NNDM algorithm uses an LLO strategy, it is computationally intensive and cannot be used with large datasets. To overcome this issue, Linnenbrink et al. (2023) suggested a k-fold variant called the kNNDM algorithm.

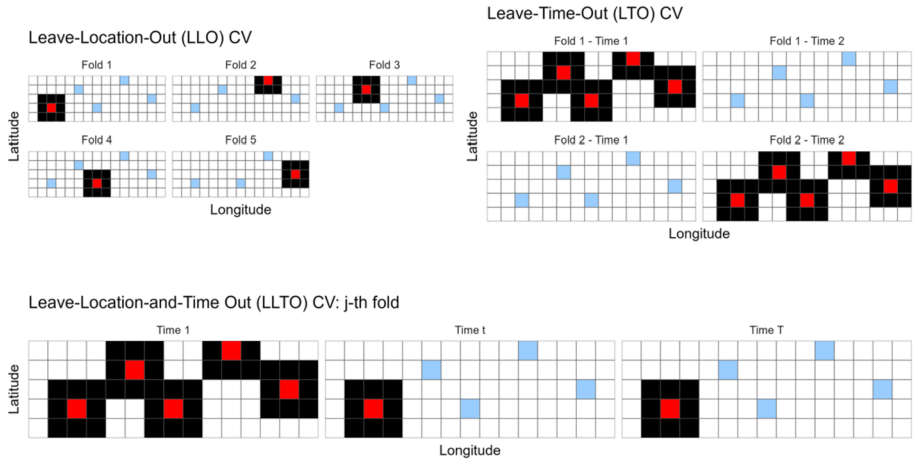

FIG 6. *Schematic spatial representation of the spatiotemporal CV schemes. In LTO, the algorithm iterates over rows; in LLO, it iterates over columns; in LLTO, it iterates over cells. Blue cells represent values used for training the model, while red cells represent points in the test set. Black blocks are observations within the spatiotemporal buffer that separate the test set from the training set.*

6. Software

In this section, we provide a concise yet comprehensive overview of available software and code for implementing penalised regression models, along with the cross-validation strategies discussed above. We primarily focus on two major programming languages for statistical data analysis, that is, `R` (R Core Team, 2023) and `Matlab` (Inc., 2022) while also including references to other frameworks where relevant. The following section encompasses not only geostatistical modelling packages but also those for generalised linear models (GLMs) and spatial regression.

6.1. Software for penalised spatiotemporal estimation

At first glance, the implementation of penalised regression algorithms presents a highly fragmented landscape. Numerous well-documented and structured libraries are available for regularised estimation across the most common programming languages for cross-sectional or panel data. In contrast, code is typically provided in an unstructured manner for penalised geostatistical models. Authors often release it in publicly accessible folders without integrating it into formal libraries, partially with minimal documentation. Nevertheless, some spatiotemporal applications can directly be implemented in available software packages for panel or cross-sectional data.

In the `R` programming language, the state-of-the-art library for penalised regression is `glmnet` (Friedman, Hastie and Tibshirani, 2010; Simon et al., 2011)[3]. `glmnet` supports a broad range of regression models with various LASSO-like penalised terms for dimension reduction and prediction in GLMs, and it includes native cross-validation functions for all supported models. For multivariate and group-structured settings, extensions such as graphical LASSO and (sparse) group LASSO are available in the `glasso` (Friedman, Hastie and Tibshirani, 2019), `cglasso` (Augugliaro et al., 2023), and `sparsegl` (Liang et al., 2024) packages. For a comprehensive review of other `R` packages related to penalised regression, we refer readers to Tay, Narasimhan and Hastie (2023).

For time-series analysis, several libraries stand out, including `LasForecast` (Lee, Shi and Gao, 2022), `midasml` (Babii, Ghysels and Striaukas, 2022), and `BigVAR` (Nicholson, Matteson and Bien, 2017a,b). `LasForecast` provides a framework for high-dimensional time series forecasting, with a focus on linear models, automatic parameter tuning, and cross-validation via rolling window forecasting. `midasml` implements sparse group LASSO penalisation for estimating and forecasting high-dimensional mixed-frequency time-series and panel data regression models. Finally, `BigVAR` handles high-dimensional multivariate time series by applying structured penalties to vector autoregressive models, making it particularly useful for economic and environmental forecasting.

For spatiotemporal data, a notable contribution comes from the `SpTe2M` library (Yang and Qiu, 2024). This software applies an exponentially weighted spatial LASSO penalisation (Samarov, Hwang and Litorja, 2015) to implement spatiotemporal process monitoring and detect potential change points over time. Additionally, several libraries support other regularisation techniques, such as penalised splines, which are frequently adapted for spatial and spatiotemporal frameworks. For example, the `R` package `pspatreg` (Basile et al., 2014; Mínguez, Basile and Durbán, 2020) employs penalised splines to estimate static and dynamic geoadditive semiparametric spatial regression models with spatial lags (e.g., SAR or spatial error models). Similarly, the `mgcv` package (Wood, 2017) implements generalised ridge regression with multiple smoothing parameters, making it suitable for fitting generalised additive mixed models with extensions to spatiotemporal, panel, and grouped data.

Similar to `R`, the `Matlab` language also provides several functions for penalised regression in both frequentist (McIlhagga, 2016, see, for example, the `penalized` toolbox for penalised GLMs with customisable optimisation settings) and Bayesian frameworks (Makalic and Schmidt, 2016, for instance, the `BayesReg` toolbox). A notable example of a spatiotemporal modelling toolbox is `DSTEM` (Finazzi and Fassò, 2015; Wang, Finazzi and Fassò, 2021), which supports the fitting, mapping, and validation of mixed models for univariate, multivariate, and functional spatiotemporal data using an efficient state space representation. In terms of dimension reduction through penalised maximum likelihood estimation, as discussed in Section 3.1, Fassò, Maranzano and Otto (2022) and

[3]As of October 2024, using the `dlstats` library (Yu, 2023), `glmnet` has been downloaded nearly 6 million times since January 2019, with a steady increase over time.

Maranzano, Otto and Fassò (2023) introduced new regularisation frameworks for multivariate and functional data within the DSTEM environment.

6.2. Software for spatiotemporal cross-validation

Regarding the task of model validation and selection through cross-validation, several solutions for spatial and spatiotemporal CV methods are implemented in the two considered statistical software.

Within the `R` programming language, the following libraries are available:

1. `sperrorest` (Brenning, 2012) implements distance-based K-means spatial partitioning,
2. `blockCV` (Valavi et al., 2019) 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 an interactive tool for visualising spatial blocks as a function of folds and block/buffer sizes,
3. `CAST` (Meyer, Milà and Ludwig, 2022; Meyer et al., 2024) implements several spatiotemporal partitioning strategies, including NNDM, kNNDMCV, LLO, LTO and LLTO schemes, as well as allows performing spatial variable selection to select suitable predictor variables according to their contribution to the spatial model performance,
4. `mlr3spatiotempcv` (Schratz et al., 2021) is a package being part of the `mlr3` ecosystem (Lang et al., 2019) which provides a unified implementation of a wide range of statistical learning 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`.

Moving to the `Matlab` environment, available software for spatiotemporal data partitioning includes the two-fold spatial CV strategy implemented in the DSTEM package (Wang, Finazzi and Fassò, 2021) for the so-called hidden dynamic geostatistical model. Furthermore, spatiotemporal application of target-oriented CV schemes can be found in Otto et al. (2024) and Maranzano, Otto and Fassò (2023).

7. Summary and conclusion

In the field of spatiotemporal statistics and econometrics, this review has demonstrated the vital role of penalised methods in coping with the growing complexity of modern spatiotemporal data. With the increasing availability of georeferenced 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, dimension 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 a 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. Whereas the literature on regularised estimation for spatiotemporal models is dominated by Gaussian models, there are some attempts explicitly designed for non-Gaussian data, which warrants a further notable avenue for future research.

There is a need for a more extensive exploration of regularised methods in spatial autoregression/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 spatiotemporal 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.

Acknowledgments

We sincerely thank the editors and anonymous reviewers for their valuable recommendations and suggestions, which helped improve the paper.

Funding

The work is partially funded by

- the project "Study of mobile phone siGNals for the evalUation of the interconnections between Mobility and the environment in Lombardia (SIGNUM)" CUP: F53D23010910001- PRIN 2022 PNRR M4C2 – financed by the European Union – Next Generation EU (DD MUR n. 1409 del 14/09/2022)
- National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3 – Call for tender No. 341 of 15/03/2022 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU. Award Number: PE_00000018, Concession Decree No. 1558 of 11/10/2022 adopted by the Italian Ministry of University and Research, CUP F83C22001720001, GROWING RESILIENT INCLUSIVE AND SUSTAINABLE – GRINS
- DFG individual grant 501539976 "Statistical Learning of High-Dimensional Spatial Dependence Structures"

References

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

Akaike, H. (1973). *Information Theory and an Extension of the Maximum Likelihood Principle*. Akadémiai Kiadó, Budapest, Hungary. MR0483125

Al-Momani, M., Hussein, A. A. and Ahmed, S. E. (2017). Penalty and related estimation strategies in the spatial error model. *Statistica Neerlandica* **71** 4–30. MR3605362

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** 327–342. https://doi.org/10.1111/jtsa.12442. MR3946156

Anselin, L. (2010). Thirty years of spatial econometrics. *Papers in Regional Science* **89** 3–26.

Arbia, G. (2016). Spatial econometrics: A rapidly evolving discipline. *Econometrics* **4** 18.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys* **4** 40–79. MR2602303

Aswi, A., Cramb, S., Moraga, P. and Mengersen, K. (2019). Bayesian spatial and spatio-temporal approaches to modelling Dengue Fever: a systematic review. *Epidemiology & Infection* **147** e33.

Augugliaro, L., Sottile, G., Wit, E. C. and Vinciotti, V. (2023). cglasso: An R package for conditional graphical LASSO inference with censored and missing values. *Journal of Statistical Software* **105** 1–58. https://doi.org/10.18637/jss.v105.i01.

Babii, A., Ghysels, E. and Striaukas, J. (2022). Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics* **40** 1094–1106. https://doi.org/10.1080/07350015.2021.1899933. MR4439275

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** 825–848. MR2523906

Basile, R., Durbán, M., Mínguez, R., María Montero, J. and Mur, J. (2014). Modeling regional economic dynamics: Spatial dependence, spatial heterogeneity and nonlinearities. *Journal of Economic Dynamics and Control* **48** 229–245. https://doi.org/10.1016/j.jedc.2014.06.011. MR3274268

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547. MR3037163

Bergmeir, C. and Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences* **191** 192–213. https://doi.org/10.1016/j.ins.2011.12.028.

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. https://doi.org/10.1016/j.csda.2014.02.001. MR3209431

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. https://doi.org/10.1016/j.csda.2017.11.003. MR3742209

Besag, J. and Green, P. J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society: Series B (Methodological)* **55** 25–37. https://doi.org/10.1111/j.2517-6161.1993.tb01467.x. MR1210422

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** 1–20. https://doi.org/10.1007/BF00116466. MR1105822

Bevilacqua, M., Fassò, A., Gaetan, C., Porcu, E. and Velandia, D. (2016). Covariance tapering for multivariate Gaussian random fields estimation. *Statistical Methods & Applications* **25** 21–37. https://doi.org/10.1007/s10260-015-0338-3. MR3460484

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

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of LASSO and Dantzig selector. *The Annals of Statistics* **37** 1705–1732, 28. MR2533469

Billé, A. G. and Arbia, G. (2019). Spatial limited dependent variable models: A review focused on specification, estimation, and health economics applica-

tions. *Journal of Economic Surveys* **33** 1531–1554.

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** 2189–2216. https://doi.org/10.1023/A:1021346712677.

Boonstra, P. S., Mukherjee, B. and Taylor, J. M. (2015). A small-sample choice of the tuning parameter in ridge regression. *Statistica Sinica* **25** 1185. MR3410304

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

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* 5372–5375. https://doi.org/10.1109/IGARSS.2012.6352393.

Brenning, A. (2022). Spatial machine-learning model diagnostics: a model-agnostic distance-based approach. *International Journal of Geographical Information Science* 1–23. https://doi.org/10.1080/13658816.2022.2131789.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media. MR2807761

Cai, L. and Maiti, T. (2020). Variable selection and estimation for high-dimensional spatial autoregressive models. *Scandinavian Journal of Statistics* **47** 587–607. https://doi.org/10.1111/sjos.12452. MR4157152

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. https://doi.org/10.1007/s13571-018-0176-z. MR4039711

Calabrese, R. and Elkink, J. A. (2014). Estimators of binary spatial autoregressive models: A Monte Carlo study. *Journal of Regional Science* **54** 664–687.

Cao, J., Guinness, J., Genton, M. G. and Katzfuss, M. (2022). Scalable Gaussian-process regression and variable selection using Vecchia approximations. *Journal of Machine Learning Research* **23** 1–30. MR4577787

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** 43–49. https://doi.org/10.1016/j.annepidem.2015.10.011.

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** 466–478. https://doi.org/10.1002/env.2410. MR3580913

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** 250–268. https://doi.org/10.1177/0962280215627298. MR3745668

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. https://doi.org/10.1016/j.csda.2022.107556. MR4452713

Cerqueira, V., Torgo, L. and Mozetič, I. (2020). Evaluating time series forecasting models: an empirical study on performance estimation methods. *Machine Learning* **109** 1997–2028. https://doi.org/10.1007/s10994-020-05910-7. MR4172968

Cerqueira, V., Torgo, L., Smailović, J. and Mozetič, I. (2017). A comparative study of performance estimation methods for time series forecasting. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* 529–538. https://doi.org/10.1109/DSAA.2017.7.

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** 117–139. MR2752029

Chattopadhyay, A. and Deb, S. (2024). A spatio-temporal model for binary data and its application in analyzing the direction of COVID-19 spread. *AStA Advances in Statistical Analysis* 1–29. MR4827900

Chavent, M., Kuentz-Simonet, V., Labenne, A. and Saracco, J. (2018). ClustGeo: an R package for hierarchical clustering with spatial constraints. *Computational Statistics* **33** 1799–1822. https://doi.org/10.1007/s00180-018-0791-1. MR3856335

Chernozhukov, V., Karl Härdle, W., Huang, C. and Wang, W. (2021). LASSO-driven inference in time and space. *The Annals of Statistics* **49** 1702–1735, 34. MR4298878

Chu, T., Zhu, J. and Wang, H. (2011). Penalized maximum likelihood estimation and variable selection in geostatistics. *The Annals of Statistics* **39** 2607–2625, 19. MR2906880

Congdon, P. (2022). A spatio-temporal autoregressive model for monitoring and predicting COVID infection rates. *Journal of Geographical Systems* **24** 583–610.

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** 209–226. MR2412639

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. MR4394913

Cressie, N., Shi, T. and Kang, E. L. (2010). Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics* **19** 724–745. MR2732500

Cressie, N. and Wikle, C. K. (2015). *Statistics for Spatio-temporal Data*. John Wiley & Sons. MR2848400

de Bruin, S., Brus, D. J., Heuvelink, G. B. M., van Ebbenhorst Tengbergen, T. and Wadoux, A. M. J. C. (2022). Dealing with clustered sam-

ples for assessing map accuracy by cross-validation. *Ecological Informatics* **69** 101665. https://doi.org/10.1016/j.ecoinf.2022.101665.

Dey, D., Datta, A. and Banerjee, S. (2022). Graphical Gaussian process models for highly multivariate spatial data. *Biometrika* **109**. MR4519112

Fahrmeir, L., Kneib, T. and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica* **14** 731–761. MR2087971

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360. MR1946581

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32** 928–961, 34. MR2065194

Fassò, A., Maranzano, P. and Otto, P. (2022). Spatiotemporal variable selection and air quality impact assessment of COVID-19 lockdown. *Spatial Statistics* **49** 100549. https://doi.org/10.1016/j.spasta.2021.100549. MR4439315

Fassò, A., Rodeschini, J., Moro, A. F., Shaboviq, Q., Maranzano, P., Cameletti, M., Finazzi, F., Golini, N., Ignaccolo, R. and Otto, P. (2023). Agrimonia: a dataset on livestock, meteorology and air quality in the Lombardy region, Italy. *Scientific Data* **10** 143. https://doi.org/10.1038/s41597-023-02034-0.

Finazzi, F. and Fassò, A. (2015). D-STEM: a software for the analysis and mapping of environmental space-time variables. *Journal of Statistical Software* **62** 1–29.

Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** 1–22. https://doi.org/10.18637/jss.v033.i01.

Friedman, J., Hastie, T. and Tibshirani, R. (2019). glasso: Graphical LASSO: Estimation of Gaussian graphical models R package version 1.11.

Furrer, R., Bachoc, F. and Du, J. (2016). Asymptotic properties of multivariate tapering for estimation and prediction. *Journal of Multivariate Analysis* **149** 177–191. MR3507322

Furrer, R., Genton, M. G. and Nychka, D. (2006a). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics* **15** 502–523. https://doi.org/10.1198/106186006X132178. MR2291261

Furrer, R., Genton, M. G. and Nychka, D. (2006b). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics* **15** 502–523. MR2291261

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. https://doi.org/10.1016/j.spasta.2015.04.001. MR3426601

Gebreab, S. Y., Duncan, D. and Kawachi, I. (2018). Statistical methods in spatial epidemiology. In *Neighborhoods and Health* 91–126. Oxford Univ.

Press.

George, E. I., Sun, D. and Ni, S. (2008). Bayesian stochastic search for VAR model restrictions. *Journal of Econometrics* **142** 553–580. https://doi.org/10.1016/j.jeconom.2007.08.017. MR2408749

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

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association* **97** 590–600. MR1941475

Gneiting, T., Kleiber, W. and Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association* **105** 1167–1177. MR2752612

Golub, G. H., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21** 215–223. https://doi.org/10.1080/00401706.1979.10489751. MR0533250

Gonella, R., Bourel, M. and Bel, L. (2022). Facing spatial massive data in science and society: variable selection for spatial models. *Spatial Statistics* 100627. https://doi.org/10.1016/j.spasta.2022.100627. MR4439338

González, J. A., Rodríguez-Cortés, F. J., Cronie, O. and Mateu, J. (2016). Spatio-temporal point process statistics: A review. *Spatial Statistics* **18** 505–544. MR3575505

Hastie, T. (2020). Ridge regularization: an essential concept in data science. *Technometrics* **62** 426–433. https://doi.org/10.1080/00401706.2020.1791959. MR4165989

Hastie, T., Tibshirani, R. and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* **2**. Springer. MR2722294

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* 425–433.

Hewamalage, H., Ackermann, K. and Bergmeir, C. (2023). Forecast evaluation for data scientists: Common pitfalls and best practices. *Data Mining and Knowledge Discovery* **37** 788–832. https://doi.org/10.1007/s10618-022-00894-5. MR4552567

Hjorth, U. (1982). Model selection and forward validation. *Scandinavian Journal of Statistics* **9** 95–105. MR0668685

Hofierka, J., Parajka, J., Mitasova, H. and Mitas, L. (2002). Multivariate interpolation of precipitation using regularized spline with tension. *Transactions in GIS* **6** 135–150. https://doi.org/10.1111/1467-9671.00101.

Hsiang, T. C. (1975). A Bayesian view on ridge regression. *Journal of the Royal Statistical Society. Series D (The Statistician)* **24** 267–268. https://doi.org/10.2307/2987923.

Hsu, N.-J., Chang, Y.-M. and Huang, H.-C. (2012). A group LASSO approach for non-stationary spatial-temporal covariance estimation. *Environmetrics* **23** 12–23. https://doi.org/10.1002/env.1130. MR2873779

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** 963–983. MR2791264

Inc., T. M. (2022). MATLAB version: 9.13.0 (R2022b), Natick, Massachusetts, United States.

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** 497–510. https://doi.org/10.1016/0304-3932(94)90030-2.

Jiang, G. and Wang, W. (2017). Markov cross-validation for time series model evaluations. *Information Sciences* **375** 219–233. https://doi.org/10.1016/j.ins.2016.09.061.

Kang, M. and Katzfuss, M. (2023). Correlation-based sparse inverse Cholesky factorization for fast Gaussian-process inference. *Statistics and Computing* **33** 56. MR4568225

Katzfuss, M. and Cressie, N. (2012). Bayesian hierarchical spatio-temporal smoothing for very large datasets. *Environmetrics* **23** 94–107. https://doi.org/10.1002/env.1147. MR2873787

Katzfuss, M. and Schäfer, F. (2024). Scalable Bayesian transport maps for high-dimensional non-Gaussian spatial fields. *Journal of the American Statistical Association* **119** 1409–1423. MR4765998

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** 1545–1555. MR2504203

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

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

Krock, M., Kleiber, W. and Becker, S. (2021). Nonstationary modeling with sparsity for spatial data via the basis graphical LASSO. *Journal of Computational and Graphical Statistics* **30** 375–389. MR4270511

Krock, M. L., Kleiber, W., Hammerling, D. and Becker, S. (2023). Modeling massive highly multivariate nonstationary spatial data with the basis graphical LASSO. *Journal of Computational and Graphical Statistics* **32** 1472–1487. MR4669262

Lam, C. and Souza, P. C. (2016). Detection and estimation of block structure in spatial weight matrix. *Econometric Reviews* **35** 1347–1376. MR3511024

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** 693–710. MR4115427

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** 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** e2465. https://doi.org/10.1002/env.2465. MR3733023

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J. and Bretagnolle, V. (2014). Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography* **23** 811–820. https://doi.org/10.1111/geb.12161.

Lee, D. (2011). A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and spatio-temporal epidemiology* **2** 79–89.

Lee, J. H., Shi, Z. and Gao, Z. (2022). On LASSO for predictive regression. *Journal of Econometrics* **229** 322–349. https://doi.org/10.1016/j.jeconom.2021.02.002. MR4429568

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* **9** 699. MR3256061

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** 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** 410–423. https://doi.org/10.1198/016214507000001337. MR2420243

Liang, X., Cohen, A., Sólon Heinsfeld, A., Pestilli, F. and McDonald, D. J. (2024). sparsegl: An R package for estimating sparse group LASSO. *Journal of Statistical Software* **110** 1–23. https://doi.org/10.18637/jss.v110.i06.

Lichstein, J. W., Simons, T. R., Shriner, S. A. and Franzreb, K. E. (2002). Spatial autocorrelation and autoregressive models in ecology. *Ecological Monographs* **72** 445–463.

Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44** 187–192. https://doi.org/10.2307/2333251.

Linnenbrink, J., Milà, C., Ludwig, M. and Meyer, H. (2023). kNNDM: k-fold nearest neighbour distance matching cross-validation for map accuracy estimation. *EGUsphere* **2023** 1–16. https://doi.org/10.5194/egusphere-2023-1308.

Liu, T. (2017). Model selection and adaptive LASSO estimation of spatial models [Doctoral dissertation]. *Ohio State University* **http://rave.ohiolink.edu/etdc/view?acc_num=osu1500379101560737**.

Liu, Y. (2022). Adaptive LASSO variable selection method for semiparametric spatial autoregressive panel data model with random effects. *Communications in Statistics – Theory and Methods* 1–19. https://doi.org/10.1080/03610926.2022.2119088. MR4698128

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. MR3809257

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. https://doi.org/10.1016/j.jag.2016.03.003.

Makalic, E. and Schmidt, D. F. (2016). High-dimensional Bayesian regularised regression with the BayesReg package. *arXiv preprint arXiv:1611.06649*. https://doi.org/10.48550/arXiv.1611.06649.

Manski, C. F. (1993). Identification of endogenous social effects: the reflection problem. *The Review of Economic Studies* **60** 531–542. MR1236836

Maranzano, P., Otto, P. and Fassò, A. (2023). Adaptive LASSO estimation for functional hidden dynamic geostatistical model. *Stochastic Environmental Research and Risk Assessment* **37** 3615–3637. https://doi.org/10.1007/s00477-023-02466-5.

McIlhagga, W. (2016). penalized: A MATLAB toolbox for fitting generalized linear models with penalties. *Journal of Statistical Software* **72** 1–21. https://doi.org/10.18637/jss.v072.i06.

Merk, M. S. and Otto, P. (2021). Directional spatial autoregressive dependence in the conditional first-and second-order moments. *Spatial Statistics* **41** 100490. MR4198095

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** e2705. MR4376814

Meyer, H., Milà, C. and Ludwig, M. (2022). CAST: 'caret' applications for spatial-temporal models. *CRAN Package, https://CRAN.R-project.org/package=CAST*. https://doi.org/10.32614/CRAN.package.CAST.

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** 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** 2208. https://doi.org/10.1038/s41467-022-29838-9.

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. https://doi.org/10.1016/j.ecolmodel.2019.108815.

Meyer, H., Ludwig, M., Milà, C., Linnenbrink, J. and Schumacher, F. (2024). The CAST package for training and assessment of spatial prediction models in R. *arXiv preprint arXiv:2404.06978*.

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* **13** 1304–1316. https://doi.org/10.1111/2041-210X.13851.

Mínguez, R., Basile, R. and Durbán, M. (2020). An alternative semiparametric model for spatial panel data. *Statistical Methods & Applications* **29** 669–708. https://doi.org/10.1007/s10260-019-00492-8. MR4174682

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83** 1023–1032. https://doi.org/10.1080/01621459.1988.10478694. MR0997578

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** 779–800. https://doi.org/10.1111/rssb.12195. MR3641407

Nicholson, W. B., Matteson, D. S. and Bien, J. (2017a). VARX-L: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting* **33** 627–651. https://doi.org/10.1016/j.ijforecast.2017.01.003.

Nicholson, W., Matteson, D. and Bien, J. (2017b). BigVAR: Tools for modeling sparse high-dimensional multivariate time series. *arXiv preprint arXiv:1702.07094*.

Nychka, D., Wikle, C. and Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling* **2** 315–331. MR1951588

Otto, P. and Steinert, R. (2023). Estimation of the spatial weighting matrix for spatiotemporal data under the presence of structural breaks. *Journal of Computational and Graphical Statistics* **32** 696–711. MR4592940

Otto, P., Fusta Moro, A., Rodeschini, J., Shaboviq, Q., Ignaccolo, R., Golini, N., Cameletti, M., Maranzano, P., Finazzi, F. and Fassò, A. (2024). Spatiotemporal modelling of $PM_{2.5}$ concentrations in Lombardy (Italy): a comparative study. *Environmental and Ecological Statistics* **31** 245–272. https://doi.org/10.1007/s10651-023-00589-0.

Paciorek, C. J. (2007). Computational techniques for spatial logistic regression with large data sets. *Computational Statistics & Data Analysis* **51** 3631–3653. MR2364480

Pannu, J. and Billor, N. (2017). Robust group-LASSO for functional regression model. *Communications in Statistics – Simulation and Computation* **46** 3356–3374. https://doi.org/10.1080/03610918.2015.1096375. MR3656106

Park, T. and Casella, G. (2008). The Bayesian LASSO. *Journal of the American Statistical Association* **103** 681–686. https://doi.org/10.1198/016214508000000337. MR2524001

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 map-

ping models. *Nature Communications* **11** 4540. https://doi.org/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. *International Journal of Geographical Information Science* **31** 2001–2019. https://doi.org/10.1080/13658816.2017.1346255.

Polson, N. G. and Scott, J. G. (2012). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis* **7** 887–902, 16. MR3000018

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** 888–898. MR3538713

Porcu, E., Furrer, R. and Nychka, D. (2021). 30 Years of space–time covariance functions. *Wiley Interdisciplinary Reviews: Computational Statistics* **13** e1512. MR4218945

Porcu, E. and White, P. A. (2022). Random fields on the hypertorus: covariance modeling and applications. *Environmetrics* **33** e2701. https://doi.org/10.1002/env.2701. MR4376811

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics* **99** 39–61. https://doi.org/10.1016/S0304-4076(00)00030-0.

Reyes, P. E., Zhu, J. and Aukema, B. H. (2012). 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. https://doi.org/10.1007/s13253-012-0103-0. MR2993279

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** 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. MR3634300

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** 1138–1148. https://doi.org/10.1016/j.ijforecast.2018.10.001.

Samarov, D. V., Hwang, J. and Litorja, M. (2015). The spatial LASSO with applications to unmixing hyperspectral biomedical images. *Technometrics* **57** 503–513. https://doi.org/10.1080/00401706.2014.979950. MR3425487

Schäfer, F., Katzfuss, M. and Owhadi, H. (2021). Sparse Cholesky factorization by Kullback–Leibler minimization. *SIAM Journal on Scientific Computing* **43** A2019–A2046. MR4267493

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. https://doi.org/10.1016/j.ecolmodel.2019.06.002.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464. MR0468014

Silverman, B. and Ramsay, J. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer New York, NY. https://doi.org/10.1007/b98886. MR1910407

Simon, N. and Tibshirani, R. (2012). Standardization and the group LASSO penalty. *Statistica Sinica* **22** 983. MR2987480

Simon, N., Friedman, J. H., Hastie, T. and Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software* **39** 1–13. https://doi.org/10.18637/jss.v039.i05.

Smart, S., Henrys, P., Scott, W., Hall, J., Evans, C., Crowe, A., Rowe, E., Dragosits, U., Page, T., Whyatt, J. et al. (2010). Impacts of pollution and climate change on ombrotrophic Sphagnum species in the UK: analysis of uncertainties in two empirical niche models. *Climate Research* **45** 163–176.

Stein, M. L. (2005). Space–time covariance functions. *Journal of the American Statistical Association* **100** 310–321. MR2156840

Stein, M. L. (2013). Statistical properties of covariance tapers. *Journal of Computational and Graphical Statistics* **22** 866–885. MR3173747

Stein, M. L. (2014). Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statistics* **8** 1–19. https://doi.org/10.1016/j.spasta.2013.06.003. MR3326818

Stein, M. L., Chi, Z. and Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66** 275–296. MR2062376

Sugasawa, S. and Murakami, D. (2021). Spatially clustered regression. *Spatial Statistics* **44** 100525. https://doi.org/10.1016/j.spasta.2021.100525. MR4277047

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting* **16** 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0.

Tay, J. K., Narasimhan, B. and Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software* **106** 1–31. https://doi.org/10.18637/jss.v106.i01.

R Core Team (2023). R: A language and environment for statistical computing R Foundation for Statistical Computing, Vienna, Austria.

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** 2173–2179. https://doi.org/10.1016/j.quascirev.2005.05.001.

Telford, R. J. and Birks, H. J. B. (2009). Evaluation of transfer functions in spatially structured environments. *Quaternary Science Reviews* **28** 1309–1316. https://doi.org/10.1016/j.quascirev.2008.12.020.

Thomas-Agnan, C., Laurent, T., Ruiz-Gazen, A., Nguyen, T. H. A., Chakir, R. and Lungarska, A. (2021). Spatial simultaneous autoregressive models for compositional data: Application to land use. In *Advances in Compositional Data Analysis: Festschrift in Honour of Vera Pawlowsky-Glahn* 225–249. Springer. MR4299318

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)* **58** 267–288. MR1379242

Trachsel, M. and Telford, R. J. (2016). Technical note: Estimating unbiased transfer-function performances in spatially structured environments. *Climate of the Past* **12** 1215–1223. https://doi.org/10.5194/cp-12-1215-2016.

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** 225–232. https://doi.org/10.1111/2041-210X.13107.

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** 2–25. https://doi.org/10.1111/sjos.12335. MR3915265

van Erp, S., Oberski, D. L. and Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology* **89** 31–50. https://doi.org/10.1016/j.jmp.2018.12.004. MR3903921

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** 297–312. MR0964183

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. MR3809256

Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M. and Fortin, M.-J. (2018). Spatial autoregressive models for statistical inference from ecological data. *Ecological Monographs* **88** 36–59.

Wallin, J. and Bolin, D. (2015). Geostatistical modelling using non-Gaussian Matérn fields. *Scandinavian Journal of Statistics* **42** 872–890. MR3391697

Wang, Y., Finazzi, F. and Fassò, A. (2021). D-STEM v2: a software for modeling functional spatio-temporal data. *Journal of Statistical Software* **99** 1–29. https://doi.org/10.18637/jss.v099.i10.

Wang, H. and Zhu, J. (2009). Variable selection in spatial regression via penalized least squares. *Canadian Journal of Statistics* **37** 607–624. https://doi.org/10.1002/cjs.10032. MR2588952

Wen, Y., Shen, X. and Lu, Q. (2018). Genetic risk prediction using a spatial autoregressive model with adaptive LASSO. *Statistics in Medicine* **37** 3764–3775. https://doi.org/10.1002/sim.7832. MR3869153

Wiemann, P. F. and Katzfuss, M. (2023). Bayesian nonparametric generative modeling of large multivariate non-Gaussian spatial fields. *Journal of Agricultural, Biological and Environmental Statistics* **28** 597–617. MR4675166

Wikle, C. K., Berliner, L. M. and Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics* **5** 117–154. https://doi.org/10.1023/A:1009662704779.

Wikle, C. K., Zammit-Mangion, A. and Cressie, N. (2019). *Spatio-temporal Statistics with R*. Chapman and Hall/CRC.

Wolpert, D. H. and Strauss, C. E. M. (1996). *What Bayes has to Say About the Evidence Procedure*. Springer Netherlands, Dordrecht. https://doi.org/10.1007/978-94-015-8729-7_3.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*, 2nd ed. Chapman and Hall/CRC. MR4011950

Wood, S. N. and Augustin, N. H. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling* **157** 157–177.

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** 425. https://doi.org/10.1186/s40064-016-2073-0.

Yang, K. and Qiu, P. (2024). SpTe2M: An R package for nonparametric modeling and monitoring of spatiotemporal data. *Journal of Quality Technology* **56** 140–156. https://doi.org/10.1080/00224065.2023.2278795.

Ye, J., Lazar, N. A. and Li, Y. (2011). Sparse geostatistical analysis in clustering fMRI time series. *Journal of Neuroscience Methods* **199** 336–345. MR2751536

Yu, G. (2023). dlstats: download stats of R packages R package version 0.1.7.

Yu, J., De Jong, R. and Lee, L.-f. (2008). Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large. *Journal of Econometrics* **146** 118–134. MR2459648

Zapata, J., Oh, S.-Y. and Petersen, A. (2022). Partial separability and functional graphical models for multivariate Gaussian processes. *Biometrika* **109** 665–681. MR4472841

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques*. MR0881437

Zeng, X. and Martinez, T. R. (2000). Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence* **12** 1–12. https://doi.org/10.1080/095281300146272.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894–942. https://doi.org/10.

1214/09-AOS729. MR2604701

Zhang, H., Huang, X., Gan, J., Karmaus, W. and Sabo-Attwood, T. (2016). A two-component G-prior for variable selection. *Bayesian Analysis* **11** 353–380, 28. MR3471994

Zhao, Y. and Karypis, G. (2002). Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the 11th International Conference on Information and Knowledge Management* 515–524.

Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. *The Journal of Machine Learning Research* **7** 2541–2563. MR2274449

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** 389–402. https://doi.org/10.1111/j.1467-9868.2010.00739.x. MR2758118

Zhu, Z. and Liu, Y. (2009). Estimating spatial covariance using penalised likelihood with weighted L_1 penalty. *Journal of Nonparametric Statistics* **21** 925–942. MR2572592

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association* **101** 1418–1429. MR2279469

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* **36** 1509–1533. https://doi.org/10.1214/009053607000000802. MR2435443