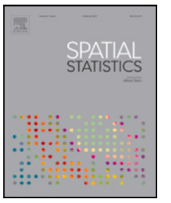
Spatial Statistics 62 (2024) 100843

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta



Automatic cross-validation in structured models: Is it time to leave out leave-one-out?

Aritz Adin [*,1], Elias Teixeira Krainski[2], Amanda Lenzi[3], Zhedong Liu[4], Joaquín Martínez-Minaya[5], Håvard Rue[2]

Universidad Pública de Navarra, Campus Arrosadía, 31006 Pamplona, Spain

ARTICLE INFO

Keywords:
Cross-validation
Hierarchical models
INLA
Spatial statistics

ABSTRACT

Standard techniques such as leave-one-out cross-validation (LOOCV) might not be suitable for evaluating the predictive performance of models incorporating structured random effects. In such cases, the correlation between the training and test sets could have a notable impact on the model's prediction error. To overcome this issue, an automatic group construction procedure for leave-group-out cross validation (LGOCV) has recently emerged as a valuable tool for enhancing predictive performance measurement in structured models. The purpose of this paper is (i) to compare LOOCV and LGOCV within structured models, emphasizing model selection and predictive performance, and (ii) to provide real data applications in spatial statistics using complex structured models fitted with INLA, showcasing the utility of the automatic LGOCV method. First, we briefly review the key aspects of the recently proposed LGOCV method for automatic group construction in latent Gaussian models. We also demonstrate the effectiveness of this method for selecting the model with the highest predictive performance by simulating extrapolation tasks in both temporal and spatial data analyses. Finally, we provide insights into the effectiveness of the LGOCV method in modeling complex structured data, encompassing spatio-temporal multivariate count data, spatial compositional data, and spatio-temporal geospatial data.

1. Introduction

Cross-validation is a general approach used to evaluate the predictive performance of a statistical model in the absence of new data, and it is commonly used to compute score rules for model comparison or selection. The fundamental concept underlying cross-validation is to split the observed data into a *training set* (the sample data used for estimating the model's parameters) and a *testing set* (the set of data points used to compute prediction errors based on the trained model) multiple times and use the combined prediction error to estimate the predictive accuracy of a model (Gelman et al., 1995; Hastie et al., 2009).

* Correspondence to: Department of Statistics, Computer Science and Mathematics, Public University of Navarre, Spain.
E-mail address: aritz.adin@unavarra.es (A. Adin).
[1] Department of Statistics, Computer Science and Mathematics, Institute for Advanced Materials and Mathematics (InaMat^2), Public University of Navarre, Spain.
[2] Statistics Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST).
[3] School of Mathematics, University of Edinburgh, Scotland.
[4] RIKEN Center for AI Project, Tokyo, Japan.
[5] Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València, Spain.

https://doi.org/10.1016/j.spasta.2024.100843
Received 28 November 2023; Received in revised form 5 March 2024; Accepted 3 June 2024
Available online 12 June 2024

The k-fold cross-validation (KCV) is a widely used variant of cross-validation, where data is first randomly partitioned into k equally (or nearly equally) sized mutually exclusive groups or folds, so that one fold is used to test the performance of the model fitted on the remaining $k-1$ folds. The procedure is repeated k times, with each repetition employing a distinct subset of the observed data as the validation set. However, a poorly chosen value for k may lead to a misrepresentative idea of the predictive performance of the model. For instance, it could result in a score with a high variance if the data used to fit the model give rise to highly variable test error measures. Alternatively, it could lead to a high bias, such as an overestimate of the performance of the model. Although no formal rule exists, a value of k equal to 5 or 10 is widely prevalent in the fields of statistics and applied machine learning as a trade-off between computational burden and bias reduction of the predictive error (Hastie et al., 2009; Kuhn et al., 2013).

Leave-one-out cross-validation (LOOCV) represents limiting case of KCV, wherein individual observations are systematically excluded, one at a time, to test the model's performance trained on the remaining data. That is, the data is divided into $k = n$ folds (where n is the size of the observed data) so that each data point is sequentially used as a testing set. The main advantages of LOOCV are that (i) the size of the training set closely approximates the full data set and, (ii) it provides an approximately unbiased estimate for the predictive error. While the exact computation of LOOCV is a "simple" and straightforward strategy, as we only need to fit models on all possible training sets and compute the utility function on the corresponding testing set, this method can become practically unfeasible due to the prohibitive computational cost of fitting the model as many times as the number of observations in the data. For this reason, several methods have been proposed to efficiently compute fast approximations of LOOCV. See, e.g. Bürkner et al. (2021), Held et al. (2010), Liu and Rue (2022) and references therein.

Standard cross-validation techniques like KCV and LOOCV operate under the assumption that data is sampled independently from a joint distribution, so that it can be freely split into training sets that represent the observed data and testing sets that represent the unobserved data (Arlot and Celisse, 2010). However, it is well-known that these techniques might not be suitable to measure the predictive performance of a model for correlated data, where commonly hierarchical models with structured random effects are considered to capture the temporal, spatial and/or hierarchical clustering dependence structures within the data (Roberts et al., 2017; Rabinowicz and Rosset, 2022). In such situations, the correlation existing between the training and testing sets can have a notable impact on the model's prediction error, and consequently affect the decision-making process for model selection, since standard cross-validation techniques are generally over-optimistic about the predictive performance of structured models.

A common approach for enhancing independence in cross-validation techniques is to adapt LOOCV to structured models by adjusting the training sets for each testing set, a process frequently determined by the specific demands of prediction task. For example, in the context of time series data (or other intrinsically ordered data) the prediction task could encompass various objectives such as interpolating a missing value in the observed data, predicting the response in an independent replicate of the observed time series, or forecasting the response in future time points. For the first case, using LOOCV could be a suitable strategy because assessing the model's predictive performance by removing a single data point replicates the interpolation task effectively. However, a different cross-validation scheme should be used to evaluate the model's ability when a time series model is used for forecasting purposes (extrapolation task). A detailed comparison of different cross-validation approaches for time series forecasting is described in Bergmeir and Benítez (2012).

Therefore, when handling structured models, it is essential to adapt the cross-validation design to ensure the calculation of dependable scoring rules for model comparison or selection. In practice, many real data analyses demand the use of complex models with multiple correlation structures such as random effects at different aggregation levels, making it very difficult to determine the optimal design and adapting it to each of the testing points. These problems are exacerbated when dealing with complex structures on large data sets. To address both issues concurrently, Liu and Rue (2022) introduced an automatic group construction procedure for leave-group-out cross-validation (LGOCV) to estimate the predictive performance of structured models by deriving an efficient and accurate approximation of LGOCV for latent Gaussian models within the INLA framework (Rue et al., 2009; Van Niekerk et al., 2023).

The aim of the current paper is two-fold: (i) to compare the behavior of predictive measures based on leave-one-out and leave-group-out cross-validation techniques within temporally or spatially structured models, and (ii) to provide real data applications of complex structured models fitted with INLA in the context of spatial statistics, while demonstrating the use of the automatic LGOCV method. In addition, we conduct a simulation study to evaluate the performance of the automatic group construction method for model comparison (see Supplementary Material).

The rest of the paper is structured as follows. Section 2 briefly describes the most relevant aspects of the LGOCV method for latent Gaussian models. In Section 3, we compare the extrapolation performance of several structured models using both temporal and spatial data. Our objective is to emphasize disparities between model selection driven by conventional criteria for Bayesian model comparison and predictive performance measures based on LOOCV and the recently proposed automatic LGOCV method. We also validate the models' predictive performance by imitating real extrapolation tasks. Section 4 presents three different real data applications in the field of spatial statistics. Specifically, a joint modeling of pancreatic cancer mortality and incidence data in England using spatio-temporal multivariate disease mapping models (Section 4.1), the estimation of the membership proportions to the different genetic clusters of *Arabidopsis thaliana* on the Iberian Peninsula using a Logistic Normal Dirichlet Model for spatial compositional data (Section 4.2), and modeling of wind speed data in the United Kingdom using spatio-temporal models for continuous domains with non-separable covariance structures (Section 4.3). Finally, Section 5 ends with some conclusions. The data and R code to reproduce the examples, simulation studies, and real data analysis presented in this paper are available at https://github.com/spatialstatisticsupna/INLA_groupCV.

2. Leave-group-out cross-validation for latent Gaussian models

LGOCV takes each data point, represented by y_i, as a testing set, and creates an index set I_i specific to each point based on the prediction task. Using all observed data except the data indexed by I_i, denoted by $\boldsymbol{y}_{-I_i}$, it computes the posterior predictive density $\pi(Y_i|\boldsymbol{y}_{-I_i})$ and validates the prediction using the observed data point y_i and scoring rules. However, when the model structure is complex or the prediction task is unclear, constructing I_i can be challenging. To overcome this, we can automatically construct I_i by selecting the most informative data points to predict y_i. This automatic construction is motivated by the fact that most extrapolation tasks have less information available from the data used for prediction. While constructing I_i automatically can be challenging in general model settings, it becomes straightforward when the model is represented as a latent Gaussian model. Furthermore, using a latent Gaussian model allows for fast and accurate approximation of posterior densities when excluding partial data.

A latent Gaussian model is composed of three layers that can be formulated by

$$\begin{aligned} &y_i|\eta_i, \boldsymbol{\theta} \sim \pi(y_i|\eta_i, \boldsymbol{\theta}),\\ &\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{f},\\ &\boldsymbol{f}|\boldsymbol{\theta} \sim N(0, \boldsymbol{P}_{\boldsymbol{f}}(\boldsymbol{\theta})),\\ &\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}). \end{aligned}$$

The observed data $\boldsymbol{y}$ is modeled conditionally independent given the hyperparameter $\boldsymbol{\theta}$ and its corresponding linear predictor η_i using the likelihood function $\pi(y_i|\eta_i, \theta)$. The linear predictor $\boldsymbol{\eta}$ is a linear combination of model effects $\boldsymbol{f}$, which has a multivariate Gaussian prior with a zero mean and a precision matrix $\boldsymbol{P}_{\boldsymbol{f}}(\boldsymbol{\theta})$ that depends on $\boldsymbol{\theta}$. We assign a prior distribution to $\boldsymbol{\theta}$. Using INLA method, we can have a Gaussian approximation of the posterior density $\pi(\boldsymbol{f}|\boldsymbol{\theta}, \boldsymbol{y}_{-I_i})$. The automatic groups are constructed using the correlation matrix of $\boldsymbol{f}$, which can be derived from either the prior density or the approximated posterior density at the mode of hyperparameter configuration. See Liu and Rue (2022) for further details.

We illustrate how to construct I_i using a simple example. Suppose we have correlation coefficients $\boldsymbol{C}_i$, which contain the correlation of the linear predictor η_i with all η_j, for $j = 1, \ldots, n$. If $\boldsymbol{C}_i = \{1, 1, 0.9, 0.9, 0.8, 0.8, -0.1, -0.1, 0, 0\}$, we define $\boldsymbol{L}_i$ as the sorted unique absolute values of $\boldsymbol{C}_i$, giving us $\boldsymbol{L}_i = \{1.0, 0.9, 0.8, 0.1, 0.0\}$. We then include the most correlated data points based on the m largest correlation levels in I_i. For example, if we choose $m = 3$, then $I_i = \{1, 2, 3, 4, 5, 6\}$. We refer to the parameter m as the number of level sets, which is determined by the degree of extrapolation implied by LGOCV. When the information about the degree of extrapolation is missing, an empirical suggestion is to use $m = 3$, or we could compute a sequence of m values to determine the predictive performance.

The automatic group construction procedure for LGOCV is implemented in the R-INLA package through the `inla.group.cv()` function, whose input argument is a previously fitted `inla` model. The output of this function contains: (i) the automatically constructed groups I_i by selecting the most informative data points to predict y_i based on posterior correlations between the linear predictors (which can be used as an input parameter to compute predictive measures for other models), and (ii) the cross-validated predictive density $\pi(Y_i = y_i|\boldsymbol{y}_{-I_i})$ and additional statistics to compute $E[Y_i|\boldsymbol{y}_{-I_i}]$. The efficient implementation of the automatic group construction procedure allows computations on relatively large datasets in a few seconds.

3. Evaluating cross-validation measures in structured models

In the forthcoming examples, our focus is on analyzing the extrapolation performance of structured models, in which the unobserved data is often assumed less dependent on the observed data. For this end, we compare the behavior of different models in terms of (i) Bayesian model selection criteria such as the deviance information criterion (DIC; Spiegelhalter et al., 2002) and the Watanabe–Akaike information criterion (WAIC; Watanabe, 2010), and (ii) predictive performance measures based on LOOCV and the recently proposed automatic group construction procedure for LGOCV for latent Gaussian models using the well-known INLA estimation technique. Specifically, we compute commonly used utility functions such as the logarithmic score (Gneiting and Raftery, 2007)

$$\text{LS} = -\frac{1}{n}\sum_{i=1}^{n} \log \pi(Y_i = y_i|\boldsymbol{y}_{-I_i}),$$

and the mean square prediction error

$$\text{MSPE} = \frac{1}{n}\sum_{i=1}^{n} \left(E[Y_i|\boldsymbol{y}_{-I_i}] - y_i\right)^2,$$

where $\pi(Y_i = y_i|\boldsymbol{y}_{-I_i})$ denotes the cross-validated predictive density at the observed data y_i. Note that in the case of LOOCV the training set $\boldsymbol{y}_{-I_i}$ reduces to $\mathbf{y}_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, y_n)'$. See Liu and Rue (2022) for details about the computation of $E[Y_i|\boldsymbol{y}_{-I_i}]$.

In our first example, we compare the predictive performance exhibited by different temporal models for long-term forecasting in time series data. Our second example delves into the extrapolation performance of spatial disease mapping models for areal count data, that is, predictions in non-observed spatial units. We show that DIC, WAIC and leave-one-out cross-validation (LOOCV) measures points to different models from those selected by using utility (loss) functions based on the recently proposed automatic LGOCV. We also validate the predictive performance of these models by imitating real extrapolation tasks for both examples based on temporal and spatial data. All computations are made using the recently implemented "compact" mode (Van Niekerk et al., 2023) in `R-INLA` (stable) version INLA_23.09.09 of R-4.3.1.

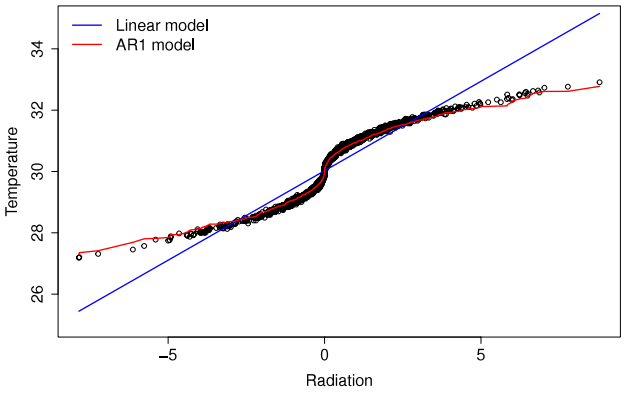


Fig. 1. Dependence structure captured between the observed shortwave radiation $\tilde{X}$ and estimated surface temperature $\hat{Y}$.

3.1. Example 1: temporal models

In this example, we want to reproduce a real situation where the aim is to perform long-term forecasting through the use of temporal models. Let us consider a scenario where our interest lies in forecasting surface temperatures using shortwave radiation as a predictor variable. We start by simulating radiation data X for a total of $n = 2000$ time points using a first-order autoregressive (AR1) model, where the autocorrelation coefficient is set at $\rho = 0.9$ and the marginal variance equals to one. Then, we generate values of surface temperature Y using the following linear model:

$$Y = \beta_0 + X + e, \quad \text{with} \quad e \sim N(0, 0.1).$$

Finally, let us assume that $\tilde{X} = X \cdot |X|$ corresponds to an imperfect observation of the shortwave radiation X (arising from potential instrumental deficiencies, for instance). The definition of $\tilde{X}$ introduces a method for establishing a non-linear relationship between the response variable (temperature) and the observed covariate, resembling the true underlying linear association with the predictor variable.

Two different models are considered to estimate the surface temperature y_t based on observed radiation data $\tilde{x}_t$ for $t = 1, \ldots, 2000$ time points.

Linear model:

$$y_t|\eta_t, \boldsymbol{\theta} \sim N(\eta_t, \tau_y),$$
$$\eta_t = \beta_0 + \beta \cdot \tilde{x}_t,$$

where β_0 is an intercept and β is the fixed effect coefficient associated to the covariate $\tilde{x}_t$.

AR1 model:

$$y_t|\eta_t, \boldsymbol{\theta} \sim N(\eta_t, \tau_y),$$
$$\eta_t = \beta_0 + u_t,$$
$$u_t = \rho \cdot u_{t-1} + \epsilon_t,$$
$$\epsilon_t \sim N(0, \tau_u)$$

where the temporally structured random effect $\mathbf{u} = (u_1, \ldots, u_{2000})'$ with AR1 prior distribution is used to model the surface temperature as a substitute of the observed predictor $\tilde{x}_t$.

Fig. 1 shows how the models capture the underlying dependence structure between the observed radiation and temperature values. Clearly, the AR1 model better captures the non-linear relationship between these variables. Posterior median estimates for the temperature values $\mathbf{y}_t = (y_1, \ldots, y_{2000})'$ obtained with both models are plotted in Fig. 2.

Table 1 displays the values of model selection criteria and predictive performance measures (logarithmic score and mean square prediction error) for LOOCV and LGOCV with automatic groups construction (m=3), derived from either the prior precision matrix or the posterior correlation matrix of each model. See Liu and Rue (2022) for further details. The groups derived from the AR1 model (as this is the true generating model) are used as a reference to make predictive measures comparable. As expected, the AR1 model is pointed out as the best model in terms of model fitting and predictive performance based on the LOOCV approach. However, predictive measures based on LGOCV points to the linear model. It should be noted that in temporally structured models, the LOOCV technique essentially evaluates their interpolation performance, i.e., how effectively the model fills in the missing values within the observed time series. In contrast, LGOCV assesses the model's capability when the prediction task is to forecast the response variable for future time points.

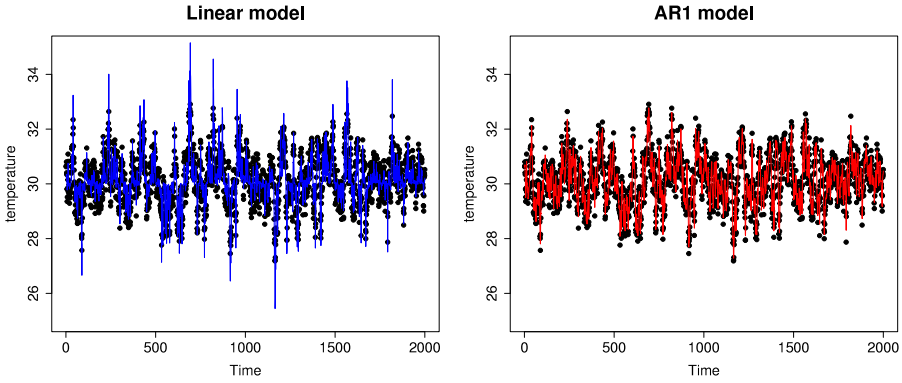


Fig. 2. True values (black dots) and posterior median estimates (colored lines) for the surface temperature $\mathbf{y}_t = (y_1, \ldots, y_{2000})'$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Model comparison for temporal data: model selection criteria (posterior mean deviance $\bar{D}$, effective number of parameters p_D, DIC and WAIC) and predictive performance measures for LOOCV and LGOCV with automatic groups construction (m=3) based on prior and posterior correlations for each temporal model.

Model	Model selection criteria				LOOCV		LGOCV			
							''prior''		''posterior''	
	$\bar{D}$	p_D	DIC	WAIC	LS	MSPE	LS	MSPE	LS	MSPE
Linear	1500.0	3.0	1503.0	1507.7	0.376	0.125	0.383	0.126	0.383	0.126
AR1	−439.5	1286.7	847.2	622.3	0.349	0.116	0.824	0.303	0.823	0.302

Table 2
Average values over 500 data sets of MAPE and RMSPE for long-term predictions.

Model	MAPE	RMSPE
Linear	0.297	0.334
AR1	0.653	0.800

To validate these measures, we conduct the following simulation study. We start by considering our training set as the time series encompassing the period $t = \{1, \ldots, 1900\}$ and compute long-term predictions for the next 100 time points using INLA. We repeat this process by shifting the time window one step backward for a total of 500 data sets. That is, our last historic data corresponds to the period $t = \{1, \ldots, 1400\}$, so that predictions are computed for future time points $\{1401, \ldots, 1500\}$. We compute the mean absolute prediction errors (MAPE) and root mean square prediction errors (RMSPE) between the posterior median estimates and true values of temperature for each data set, $s = 1, \ldots, 500$, as

$$\text{MAPE}_s = \frac{1}{100}\sum_{j=1}^{100}\left|\hat{y}_{t+j} - y_{t+j}\right| \qquad \text{and} \qquad \text{RMSPE}_s = \sqrt{\frac{1}{100}\sum_{j=1}^{100}(\hat{y}_{t+j} - y_{t+j})^2}.$$

Table 2 presents the average values of these metrics calculated across the 500 data sets. Our simulation study corroborates that the linear model outperforms the AR1 model when the objective is to perform long-term forecasting. This is something expected, as the linear model incorporates additional auxiliary information derived from the observed covariate, resulting in more accurate predictions. Conversely, the AR1 model converges toward the overall mean as time advances, as illustrated in Fig. 3.

3.2. Example 2: spatial models

The objective of this example is to validate the effectiveness of automatic groups construction within LGOCV for extrapolation prediction tasks in disease mapping models, that is, when the goal is to predict disease risks or rates in unobserved regions using spatially structured models.

We use data concerning dowry deaths across the 70 districts of Uttar Pradesh (the most populous state in India) during the year of 2011, a form of crime against women which is very specific to India. For each district, we have records of the number of observed and expected dowry deaths. The expected deaths are calculated by considering women aged 15 to 49 years as the population at risk. In addition, we also consider the following predictors as potential areal risk factors: sex ratio (number of females per 1000 males), per capita income, and number of murders per 100,000 inhabitants. In Fig. 4 we plot the maps of the standardized mortality ratio (SMR) and standardized covariates at small area level. Similar data were analyzed by Vicente et al. (2020a) and Adin et al. (2023a) to study the association between these potential risk factors and dowry deaths in Uttar Pradesh during the period 2001–2011.

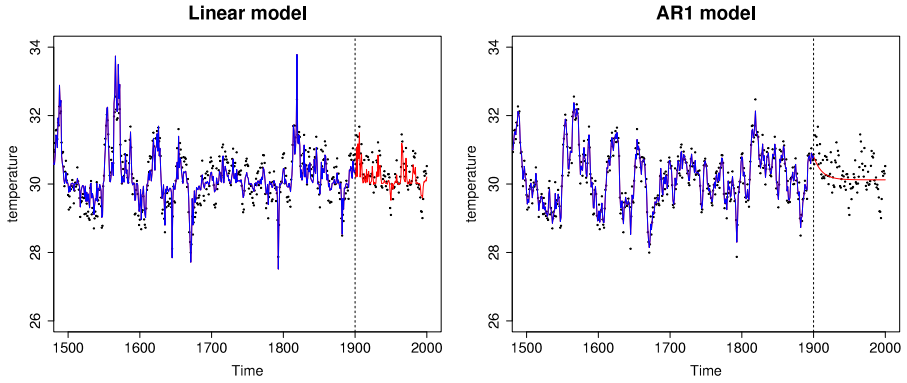


Fig. 3. True values (black dots), posterior median estimates (blue line) and forecasted values (red lines) obtained from linear and AR1 models for the first data set of the simulation study. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

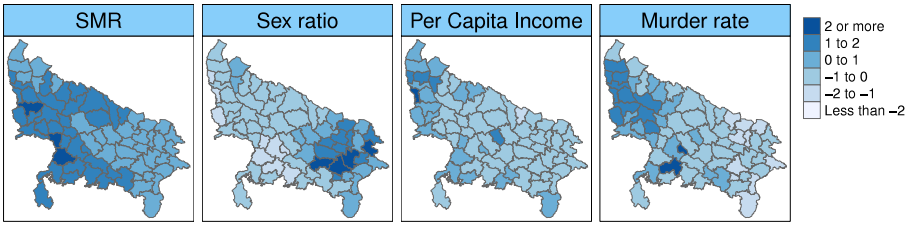


Fig. 4. Maps of the standardized mortality ratio (SMR) of dowry deaths in the districts of Uttar Pradesh and standardized potential risk factors.

Let y_i and E_i be the number of observed and expected deaths, respectively. Conditional on the relative risks r_i, the number of observed cases is assumed to follow a Poisson distribution

$$y_i|r_i, \boldsymbol{\theta} \sim \text{Poisson}(\mu_i = E_i r_i)$$
$$\log \mu_i = \log E_i + \log r_i$$

where $\log E_i$ is an offset and depending on the specification of $\eta_i = \log r_i$ different models are defined. Here, different spatially structured models are considered to estimate the relative risks of dowry deaths for each small area $i = 1, \ldots, 70$.

CAR model:

$$\eta_i = \beta_0 + \xi_i,$$

where β_0 is an intercept representing the overall log-risk and ξ_i is a spatially structured random effect for which the so-called BYM2 conditional autoregressive (CAR) prior distribution (Riebler et al., 2016) is assumed.

CAR model + covariates:

$$\eta_i = \beta_0 + x_i'\boldsymbol{\beta} + \xi_i,$$

where x_i' is the vector of standardized covariates in the ith small area and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$ is the vector of fixed effects coefficients.

P-spline model:

$$\eta_i = \beta_0 + f(s_{1i}, s_{2i}),$$

where $f(s_{1i}, s_{2i})$ is a spatial smooth function that is approximated as $f(\mathbf{s}_1, \mathbf{s}_2) \approx \mathbf{B}_s\boldsymbol{\theta}_s$, where $\mathbf{B}_s = \mathbf{B}_2 \square \mathbf{B}_1$ is the two-dimensional B-spline basis of dimension $70 \times k_s$ (with $k_s = k_1 k_2$ depending on the number of knots and the degree of the polynomials in the bases $\mathbf{B}_1$ and $\mathbf{B}_2$), obtained from the row-wise Kronecker product of the marginal bases for longitude $\mathbf{s}_1 = (s_{1,1}, \ldots, s_{1,70})'$ and latitude $\mathbf{s}_2 = (s_{2,1}, \ldots, s_{2,70})'$. To achieve smoothness, the following prior distribution is assumed for the unknown coefficients $\boldsymbol{\theta}_s = (\theta_1, \ldots, \theta_{k_s})'$

$$\boldsymbol{\theta}_s \sim N(\mathbf{0}, \mathbf{P}_s^-)$$

where $\mathbf{P}_s$ is a spatial anisotropic precision matrix that induces different amount of smoothing in longitude an latitude dimensions (see Ugarte et al., 2017). The symbol $^-$ denotes the Moore–Penrose generalized inverse of a matrix.

P-spline model + covariates:

$$\eta_i = \beta_0 + x_i'\boldsymbol{\beta} + f(s_{1i}, s_{2i})$$

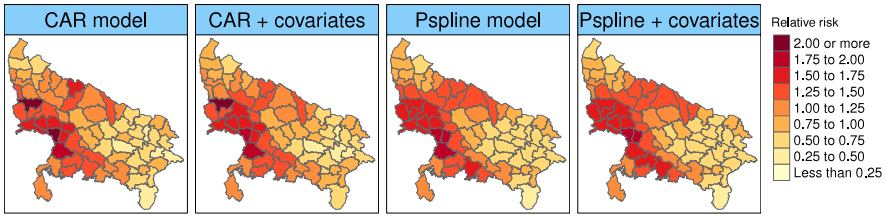


Fig. 5. Maps of posterior median estimates of dowry death relative risks obtained from the different spatial models.

Table 3
Model comparison for spatial data: model selection criteria (posterior mean deviance $\bar{D}$, effective number of parameters p_D, DIC and the WAIC) and predictive performance measures for LOOCV and LGOCV with automatic groups construction (m = 3) based on posterior correlations for each spatial model.

Model	Model selection criteria				LOOCV		LGOCV	
	$\bar{D}$	p_D	DIC	WAIC	LS	MSPE	LS	MSPE
CAR	433.2	47.5	**480.7**	**473.4**	**3.604**	**90.529**	3.611	89.480
CAR+covariates	435.4	45.6	481.1	476.8	3.616	105.042	3.614	99.982
Pspline	454.0	30.4	484.4	492.0	3.600	93.196	**3.485**	**76.146**
Pspline+covariates	459.5	30.8	490.3	501.0	3.667	101.102	3.508	82.271

Table 4
Average values over 70 data sets of MRPE and RRMSPE for predicted counts.

Model	MRPE			RRMSPE		
	k = 1	k = 2	k = 3	k = 1	k = 2	k = 3
CAR	0.273	0.325	0.370	0.354	0.442	0.521
CAR + covariates	0.281	0.299	0.310	0.357	0.395	0.410
Pspline	0.270	0.316	0.344	0.347	0.422	0.470
Pspline + covariates	0.161	0.171	0.172	0.215	0.246	0.253

The maps illustrating the posterior median estimates of $r_i = \exp(\eta_i)$ are presented in Fig. 5. In general, similar geographical distributions of mortality risks related to dowry deaths are observed across the different spatial models. Table 3 compares the values of model selection criteria and predictive performance measures considering LOOCV and LGOCV with automatic groups construction (m = 3) based on posterior correlations for each spatial model. According to DIC/WAIC measures, CAR models are preferred over P-spline models in terms of model selection criteria. In addition, predictive measures derived from LOOCV suggest that the incorporation of spatial covariates into the linear predictor of the CAR/P-spline models does not enhance their predictive performance. Considering these measure, the P-spline model with covariates exhibits the poorest performance in both the trade-off between model fit and complexity, and in terms of predictive accuracy. In contrast, the predictive metrics obtained through LGOCV indicate that the CAR models perform less effectively in predicting relative risks for unobserved small areas. This is something expected considering that CAR models induce local correlations among neighboring areas, while P-spline models are able to borrow spatial information for areas that are further apart. Similar conclusions are obtained for higher values of m.

As in the previous example, we should notice that LOOCV is not an appropriate approach to estimate the models predictive performance when the task is to predict relative risks in unobserved group of areas (extrapolation task). This concern is particularly relevant when employing spatially structured models, where the linear predictor of each area is highly correlated with those of its neighboring regions. Once again, our focus lies in validating these measures through a simulation study. For each small area, we remove its k-order neighboring or spatially contiguous areas ($k = 1, 2$ and 3) and compute posterior predictive distributions for the relative risks on those areas under the different models fitted with INLA. In Fig. 6 we show a map with the 1st, 2nd and 3rd-order neighborhood areas for a randomly selected district of Uttar Pradesh. In Table 4 we show average values of mean relative prediction errors (MRPE) and relative root mean square prediction errors (RRMSPE) computed over each data set. As expected, higher MRPE and RRMSPE values are obtained as the number of removed areas increases, which is controlled by the parameter k. Our simulation study corroborates that, as pointed out by the LGOCV approach, CAR models exhibit inferior predictive performance compared to P-spline models in the context of spatial extrapolation. In general, we observe that better predictive count estimates are obtained when including auxiliary information of area-level covariates as predictors in the models.

4. Applications

When constructing groups for different models, it is essential to choose a reference model to ensure that we can obtain comparable measures using the LGOCV approach. The simulation study presented as Supplementary Material suggests that automatic groups generated from the posterior of a suitable model fitted to a dataset is the recommended approach for evaluating various candidate

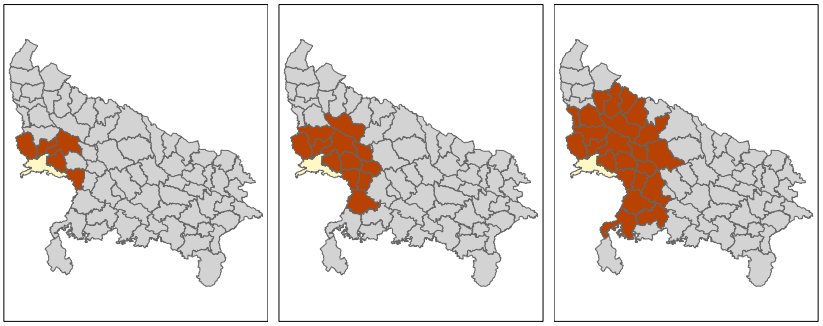


Fig. 6. Maps with 1st-order (left), 2nd-order (center) and 3rd-order (right) neighborhood areas for a selected district of Uttar Pradesh.

models. Based on these results, for the real data analyses presented in this section, we decided to use models with the lowest DIC values as the reference models for constructing the automatic groups and subsequently comparing the predictive performance of the different candidates.

The supplementary material also contains additional figures pertaining to the real data applications on spatial statistics discussed in the subsequent sections.

4.1. Joint modeling of pancreatic cancer mortality and incidence data

Bayesian hierarchical models are a powerful tool for analyzing area-level incidence or mortality data, providing valuable insight into latent spatial, temporal, and spatio-temporal patterns. These models typically rely on generalized linear mixed models that incorporate spatially and temporally structured random effects, allowing for the smoothing of disease risks or rates by incorporating information from neighboring areas and time periods. Further extensions of spatio-temporal disease mapping models aim to jointly analyze several responses by the construction of multivariate proposals based on Gaussian Markov random fields. An exhaustive review on the topic can be found in the work by MacNab (2018). Alternatively, shared-component models have been widely used to investigate closely related phenomena or diseases that share a recognized common risk factor (see Knorr-Held and Best, 2001; Held et al., 2005).

4.1.1. The models

In this section, we demonstrate the use of the automatic group construction procedure for LGOCV when the objective is to jointly model incidence and mortality data for a highly lethal disease such as pancreatic cancer using different spatio-temporal models. Specifically, we conduct an analysis using yearly count data for the male population across the 105 Clinical Commissioning Groups (CCG) of England during the period 2001–2020 (official data provided by the National Health Service in England available at https://www.cancerdata.nhs.uk/incidence_and_mortality). In these types of studies, we expect that using multivariate models will enhance the performance of individual models by leveraging the strong correlation between incidence and mortality data.

The setting for our study is the following. Let us define as y_{it1} and y_{it2} the number of mortality and incidence cases, respectively, by health-area (CCG) $i = 1, \dots, S = 105$ and time period $t = 1, \dots, T = 20$ (corresponding to the period 2001–2020). Conditional on the relative risks r_{itj} for mortality ($j = 1$) and incidence ($j = 2$) data, we assume that the number of observed cases follow a Poisson distribution with mean $\mu_{itj} = E_{itj} \cdot r_{itj}$, that is,

$$y_{itj}|r_{itj}, \boldsymbol{\theta} \sim \text{Poisson}(\mu_{itj} = E_{itj} \cdot r_{itj}),$$
$$\log \mu_{itj} = \log E_{itj} + \log r_{itj}.$$

Here E_{itj} is computed using indirect standardization as $E_{itj} = \sum_k n_{itjk} \cdot m_{jk}$, where k denotes the age-groups (< 25, $[25, 50)$, $[50, 60)$, $[60, 70)$, $[70, 80)$ and ≥ 80), n_{itjk} is the population at risk in the area i, time t and age-group k, and m_{jk} is the overall mortality/incidence rate in England during the whole study period for the kth age-group. Then, the log-risk is modeled as

$$\log r_{itj} = \alpha_j + \phi_{ij} + \gamma_{tj} + \delta_{itj}, \tag{1}$$

where α_j is an intercept representing the overall log-risk, ϕ_{ij} and γ_{tj} are the spatial and temporal main effects, and δ_{itj} is the spatio-temporal interaction for mortality and incidence data respectively.

Three different disease mapping models are considered to jointly estimate pancreatic cancer mortality and incidence data under the model formulation of Eq. (1). Firstly, independent spatio-temporal models (M1) are considered. For comparison purposes with multivariate models, if we denote as $\mathbf{r} = (\mathbf{r}_1', \mathbf{r}_2')'$ to the joined vector of mortality and incidence risks with $\mathbf{r}_j = (r_{11j}, \dots, r_{S1j}, \dots, r_{1Tj}, \dots, r_{STj})'$, we formulate the model in matrix form as

$$\begin{pmatrix} \log \mathbf{r}_1 \\ \log \mathbf{r}_2 \end{pmatrix} = (\mathbf{I}_2 \otimes \mathbf{1}_T \otimes \mathbf{1}_S) \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + (\mathbf{I}_2 \otimes \mathbf{1}_T \otimes \mathbf{I}_S) \begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \end{pmatrix} + (\mathbf{I}_2 \otimes \mathbf{I}_T \otimes \mathbf{1}_S) \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{pmatrix} + (\mathbf{I}_2 \otimes \mathbf{I}_T \otimes \mathbf{I}_S) \begin{pmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \end{pmatrix}, \tag{2}$$

Table 5
Summary of the prior distributions for the spatial, temporal and spatio-temporal random effects under the three models considered in Section 4.1.

	M1 (Independent models)	M2 (M-model)	M3 (Shared-component model)
Spatial effect	$\boldsymbol{\phi}_j \sim N(\mathbf{0}, [\tau_{\phi_j}\mathbf{R}_\phi]^-), \quad j=1,2$	$\begin{pmatrix}\boldsymbol{\phi}_1\\ \boldsymbol{\phi}_2\end{pmatrix} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_\phi^{-1}\otimes\mathbf{R}_\phi^-\right)$	$\boldsymbol{\phi} \sim N(\mathbf{0}, [\tau_\phi\mathbf{R}_\phi]^-)$
Temporal effect	$\boldsymbol{\gamma}_j \sim N(\mathbf{0}, [\tau_{\gamma_j}\mathbf{R}_\gamma]^-), \quad j=1,2$	$\begin{pmatrix}\boldsymbol{\gamma}_1\\ \boldsymbol{\gamma}_2\end{pmatrix} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_\gamma^{-1}\otimes\mathbf{R}_\gamma^-\right)$	$\boldsymbol{\gamma} \sim N(\mathbf{0}, [\tau_\gamma\mathbf{R}_\gamma]^-)$
Spatio-temporal effect	$\boldsymbol{\delta}_j \sim N(\mathbf{0}, [\tau_{\delta_j}\mathbf{R}_\delta]^-) \quad j=1,2$ Type I: $\mathbf{R}_\delta = \mathbf{I}_T\otimes\mathbf{I}_S$, Type II: $\mathbf{R}_\delta = \mathbf{R}_\gamma\otimes\mathbf{I}_S$ Type III: $\mathbf{R}_\delta = \mathbf{I}_T\otimes\mathbf{R}_\phi$, Type IV: $\mathbf{R}_\delta = \mathbf{R}_\gamma\otimes\mathbf{R}_\phi$		

where $\mathbf{1}_S$ and $\mathbf{1}_T$ are column vectors of ones of length S and T respectively, $\mathbf{I}_2$, $\mathbf{I}_S$ and $\mathbf{I}_T$ are identity matrices of dimension 2×2, $S\times S$ and $T\times T$ respectively, $\boldsymbol{\phi}_j = (\phi_{1j},\ldots,\phi_{Sj})'$ is a spatial random effect with intrinsic CAR prior distribution, $\boldsymbol{\gamma}_j = (\gamma_{1j},\ldots,\gamma_{Tj})'$ is a temporally structured random effect that follows a first order RW prior distribution, and $\boldsymbol{\delta}_j = (\delta_{11j},\ldots,\delta_{STj})'$ is a spatio-temporal random effect where the four different types of interactions originally proposed by Knorr-Held (2000) are considered. That is, for $j=1,2$

$$\boldsymbol{\phi}_j \sim N(\mathbf{0}, [\tau_{\phi_j}\mathbf{R}_\phi]^-), \qquad \boldsymbol{\gamma}_j \sim N(\mathbf{0}, [\tau_{\gamma_j}\mathbf{R}_\gamma]^-), \qquad \text{and} \qquad \boldsymbol{\delta}_j \sim N(\mathbf{0}, [\tau_{\delta_j}\mathbf{R}_\delta]^-),$$

where $\mathbf{R}_\phi$, $\mathbf{R}_\gamma$ and $\mathbf{R}_\delta$ are structure matrices for the corresponding spatial, temporal and spatio-temporal random effects respectively, and τ_{ϕ_j}, τ_{γ_j} and τ_{δ_j} are precision parameters. See, for example, Goicoa et al. (2018) for details about identifiability issues in this type of spatio-temporal models and its implementation using INLA.

Alternatively, two distinct multivariate models that enable the incorporation of correlation structures between mortality and incidence data are considered: a spatio-temporal extension of the so-called M-based model described in Vicente et al. (2020b) (M2) and models including shared-component terms for space and time similar to the ones described in Etxeberria et al. (2023) (M3). Denote as $\boldsymbol{\phi} = (\boldsymbol{\phi}_1', \boldsymbol{\phi}_2')'$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2')'$ the joined vectors of spatial and temporal random effects for mortality and incidence risks, respectively. Then, the M-model formulation (M2) is identical to the expression in Eq. (2), assuming the following prior distribution

$$\boldsymbol{\phi} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_\phi^{-1}\otimes\mathbf{R}_\phi^-\right), \qquad \text{and} \qquad \boldsymbol{\gamma} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_\gamma^{-1}\otimes\mathbf{R}_\gamma^-\right),$$

where $\boldsymbol{\Sigma}_\phi$ and $\boldsymbol{\Sigma}_\gamma$ are the covariance matrices between the spatial and temporal effects, respectively, of mortality and incidence data. See Adin et al. (2023b) for details about model implementation in INLA. Finally, we formulate our shared component model (M3) as

$$\log r_{it1} = \alpha_1 + c_\phi\phi_i + c_\gamma\gamma_t + \delta_{it1},$$

$$\log r_{it2} = \alpha_2 + \frac{1}{c_\phi}\phi_i + \frac{1}{c_\gamma}\gamma_t + \delta_{it2},$$

where ϕ_i is a shared spatial effect with intrinsic CAR prior distribution, γ_t is a shared temporal effect with first order RW prior distribution, and δ_{itj} is a specific spatio-temporal random effect to model the interaction in mortality and incidence log-risks, respectively. Finally, c_ϕ and c_γ are scaling parameters to allow a different "risk gradient" (on the log-scale) to be associated with the spatial and temporal components, respectively, for both responses. A summary with the main prior distributions for the three models considered in this section has been included in Table 5.

To fit the models, improper uniform prior distributions are given to all the standard deviations (square root inverse of precision parameters), and a Gamma(10,10) distribution is considered for the scaling parameters c_ϕ and c_γ of the shared-component model. Finally, a vague zero mean normal distribution with a precision close to zero (0.001) is given to the model intercepts.

4.1.2. Results

Table 6 presents a comparison of the different models used to estimate pancreatic cancer mortality and incidence relative risks in terms of model selection criteria and predictive performance measures. Both DIC/WAIC and predictive measures using a LOOCV approach consistently indicate that Type II interaction models should be selected. Among the considered models, the shared-component model M3 exhibits a slightly better level of performance. To compute comparable LS measures from the posterior predictive density functions $\pi(Y_{itj}|\boldsymbol{y}_{-I_{itj}})$, we use the groups I_{itj} derived from Type II interaction models with number of level sets $m=5$. As shown in Table 6, very close values are obtained for both multivariate modeling approaches M2 and M3.

An interesting aspect of the automatic group construction approach, which relies on the posterior correlation matrix of the linear predictor, is its inherent interpretability. For each testing point (area i, time t and response j), we can examine its derived groups to identify which structured effect of the model contributes to a stronger correlation in the risk estimates. In the independent model (M1), the groups assigned to all the testing points consistently correspond to adjacent years. In the M-model (M2), while the temporally structured random effect remains as the dominant pattern in the model, there are some testing points where the group of highly correlated data points also includes spatial neighbors (areas that share a common border). Finally, the groups derived from the shared-component model (M3) reveal a strong correlation between both responses (mortality and incidence data) for the majority of areas in the year 2011. This finding suggest that the shared-component model would be the most suitable choice for predicting pancreatic cancer mortality based on incidence data.

Table 6
Pancreatic cancer data: model selection criteria (DIC) and predictive performance measures (Logarithmic Score) with automatic groups construction for $m = 3, 5$, and 10.

	DIC	Logarithmic Score		
		$m = 3$	$m = 5$	$m = 10$
M1 - Type I	26 095.3	3.113	3.115	3.122
M1 - Type II	**26065.4**	**3.109**	**3.112**	**3.119**
M1 - Type III	26 120.9	3.115	3.118	3.125
M1 - Type IV	26 074.1	3.110	3.112	3.119
M2 - Type I	26 000.0	3.099	3.101	3.104
M2 - Type II	**25961.7**	**3.094**	**3.093**	**3.094**
M2 - Type III	26 018.8	3.101	3.102	3.106
M2 - Type IV	25 973.5	3.095	3.095	3.096
M3 - Type I	25 995.0	3.098	3.099	3.106
M3 - Type II	**25953.9**	**3.092**	**3.092**	**3.099**
M3 - Type III	26 012.1	3.099	3.100	3.107
M3 - Type IV	25 964.7	3.093	3.093	3.100

4.2. Spatial compositional data: the case of *Arabidopsis thaliana*

Compositional Data Analysis has gained popularity for comprehending processes wherein values correspond to D distinct categories, with their sum constituting a constant. These values typically represent proportions (or percentages) summing up to one (or 100). The data stemming from such processes are commonly referred to as Compositional Data (CoDa). Without loss of generality, we assume the constant to be one. A substantial body of literature exists, delving into the application of CoDa analysis across various domains, including Ecology (Kobal et al., 2017; Douma and Weedon, 2019), Geology (Buccianti and Grunsky, 2014; Engle and Rowan, 2014), Genomics (Tsilimigras and Fodor, 2016; Shi et al., 2016; Washburne et al., 2017; Creus Martí et al., 2022), Environmental Sciences (Aguilera et al., 2021; Mota-Bertran et al., 2022) or Medicine (Dumuid et al., 2018; Fairclough et al., 2018).

Recently, two approaches have been introduced to incorporate the CoDa analysis within the framework of latent Gaussian models. One involves utilizing Dirichlet regression (Martínez-Minaya et al., 2023), while the other proposes the Logistic Normal Dirichlet Model (LNDM) (Martínez-Minaya and Rue, 2024), which mainly uses logistic-normal distribution with Dirichlet covariance through the additive log-ratio transformation as likelihood promoting interpretation as log-ratios. In this example, we focus on the latter, illustrating the utility of an automatic group construction procedure for LGOCV in a spatial CoDa problem: the case of *Arabidopsis thaliana*.

4.2.1. The data

We conduct our study using a dataset consisting of $N = 301$ accessions of the annual plant *Arabidopsis thaliana* located on the Iberian Peninsula. For each accession, the genetic cluster (GC) membership proportions for the four genetic clusters (GC1, GC2, GC3, GC4) are available as identified in Martínez-Minaya et al. (2019). These probabilities, shown in Fig. 7, sum up to one. Our main interest lies in estimating the membership probability, which in this specific context can be interpreted as the habitat suitability for each genetic cluster. To achieve this, we employ LNDMs incorporating climate-related covariates and spatially structured effects within the linear predictor. Specifically, we employ two bioclimatic variables to define the climatic aspect: annual mean temperature (BIO1) and annual precipitation (BIO12). The dataset is available at http://dx.doi.org/10.5281/zenodo.2552025 (Martínez-Minaya et al., 2019). Prior to analysis, climate covariates are standardized.

4.2.2. The models

As mentioned earlier, four categories are employed in this problem: GC1, GC2, GC3 and GC4. So, we deal with proportions in the simplex $\mathbb{S}^4$. To construct the LNDM, the additive log-ratio transformation (alr) is required, and we select GC4 as the reference category because it has the lowest variance in its logarithm. We are thus dealing with a three dimensional normal distribution with Dirichlet covariance, $\mathcal{NDI}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The data, as well as the transformed data, are displayed in Fig. 7.

The models employed to address this issue present the subsequent structure:

$$\begin{aligned} alr(\boldsymbol{Y}) &\sim \mathcal{NDI}((\boldsymbol{\mu}^{(1)}, \dots, \boldsymbol{\mu}^{(3)}), \boldsymbol{\Sigma}), \\ \boldsymbol{\mu}^{(d)} &= \boldsymbol{X}\boldsymbol{\beta}^{(d)} + \boldsymbol{\omega}^{(d)}, \ d = 1, \dots, 3, \end{aligned} \tag{3}$$

where $alr(\boldsymbol{Y})$ represents the alr transformation applied to the compositional variable. From now on, we will refer to the transformed variable as alr-coordinates, and the vector $\boldsymbol{\mu}^{(d)} = (\mu_1^{(d)}, \dots, \mu_{301}^{(d)})'$ is the mean of the dth alr-coordinate. The design matrix $\boldsymbol{X}$ of dimension 301×3 consists of ones in its first column and climate covariate values in the remaining columns. The spatial random effect for each dth alr-coordinate is denoted by $\boldsymbol{\omega}^{(d)}$, characterized by a Matérn covariance structure. Specifically, $\boldsymbol{\omega}^{(d)}$ follows a normal distribution $\mathcal{N}(0, \boldsymbol{Q}^{-1}(\sigma_{\boldsymbol{\omega}}, \phi))$, where $\sigma_{\boldsymbol{\omega}}$ represents the standard deviation of the spatial effect, and ϕ its range. The parameter vector $\boldsymbol{\beta} = (\beta_0^{(d)}, \beta_1^{(d)}, \beta_2^{(d)})'$ corresponds to fixed effects. The latent field comprises both the fixed effects parameters and realizations of the random field. This configuration can be succinctly presented as:

$$\boldsymbol{\mathcal{X}} = \{\boldsymbol{\beta}^{(d)}, \boldsymbol{\omega}^{(d)} : d = 1, \dots, 3\}.$$

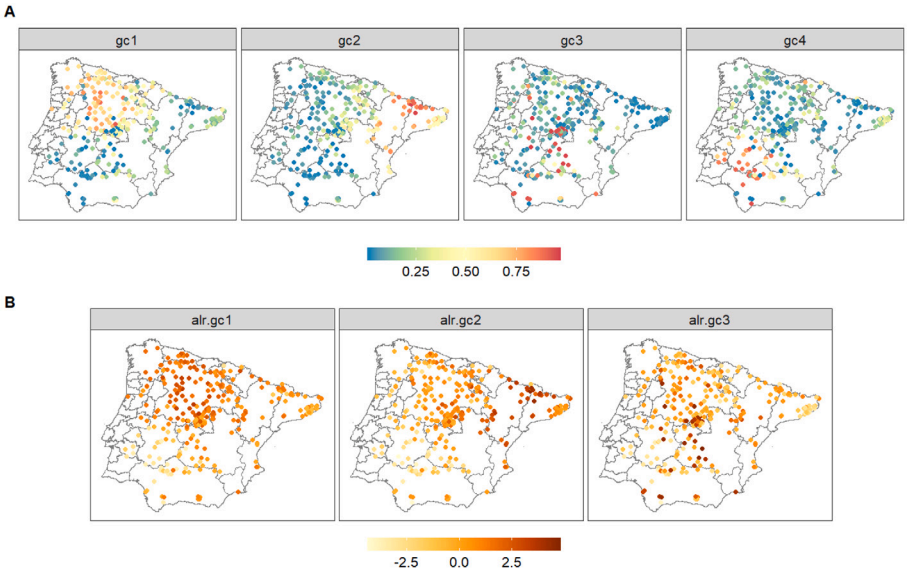


Fig. 7. CoDa of 301 accessions of the annual plant *Arabidopsis thaliana* located on the Iberian Peninsula. **A**: Membership proportion to each of the 4 genetic clusters extracted from Martínez-Minaya et al. (2019). **B**: *alr*-coordinates generated from the *alr* transformation using GC4 as reference.

Table 7
Different structures included in the model in an additive way with their corresponding latent field and the set hyperparameters to be estimated.

Models	*Predictor*	*Latent field* ($\boldsymbol{\mathcal{X}}$)	*Hyperparameters* ($\boldsymbol{\theta}$)
Type I	$\boldsymbol{X\beta}$	$\{\beta_0, \beta_1, \beta_2\}$	$\{\sigma_d^2, \gamma\}$
Type II	$\boldsymbol{X\beta}^{(d)}$	$\{\beta_0^{(d)}, \beta_1^{(d)}, \beta_2^{(d)}\}$	$\{\sigma_d^2, \gamma\}$
Type III	$\boldsymbol{X\beta} + \boldsymbol{\omega}$	$\{\beta_0, \beta_1, \beta_2, \omega_1, \ldots, \omega_N\}$	$\{\sigma_d^2, \gamma, \sigma_\omega, \phi\}$
Type IV	$\boldsymbol{X\beta}^{(d)} + \boldsymbol{\omega}$	$\{\beta_0^{(d)}, \beta_1^{(d)}, \beta_2^{(d)}, \omega_1, \ldots, \omega_N\}$	$\{\sigma_d^2, \gamma, \sigma_\omega, \phi\}$
Type V	$\boldsymbol{X\beta} + \boldsymbol{\omega}^{*(d)}$	$\{\beta_0, \beta_1, \beta_2, \omega_1, \ldots, \omega_N\}$	$\{\sigma_d^2, \gamma, \sigma_\omega, \phi, \alpha^{(1)}, \alpha^{(2)}\}$
Type VI	$\boldsymbol{X\beta}^{(d)} + \boldsymbol{\omega}^{*(d)}$	$\{\beta_0^{(d)}, \beta_1^{(d)}, \beta_2^{(d)}, \omega_1, \ldots, \omega_N\}$	$\{\sigma_d^2, \gamma, \sigma_\omega, \phi, \alpha^{(1)}, \alpha^{(2)}\}$
Type VII	$\boldsymbol{X\beta} + \boldsymbol{\omega}^{(d)}$	$\{\beta_0, \beta_1, \beta_2, \omega_1^{(d)}, \ldots, \omega_N^{(d)}\}$	$\{\sigma_d^2, \gamma, \sigma_\omega, \phi\}$
Type VIII	$\boldsymbol{X\beta}^{(d)} + \boldsymbol{\omega}^{(d)}$	$\{\beta_0^{(d)}, \beta_1^{(d)}, \beta_2^{(d)}, \omega_1^{(d)}, \ldots, \omega_N^{(d)}\}$	$\{\sigma_d^2, \gamma, \sigma_\omega, \phi\}$

In contrast, the vector $\boldsymbol{\theta}_1 = \{\sigma_d^2, \gamma : d = 1, \ldots, 3\}$ encompasses hyperparameters linked to the likelihood. On the other hand, $\boldsymbol{\theta}_2 = \{\sigma_\omega, \phi\}$ comprises hyperparameters pertaining to the spatial random effect. These combined elements collectively form the set of hyperparameters. Gaussian prior distributions are assumed for the fixed effects, while PC-priors are employed for the hyperparameters, following Simpson et al. (2017). Based on the model structure defined in Eq. (3) where a spatial effect is included, we focus on the eight structures presented in Martínez-Minaya and Rue (2024):

- Type I: share the same parameters for fixed effects, and do not include spatial random effects.
- Type II: have different parameters for fixed effects, and do not include spatial random effects.
- Type III: share the same parameters for fixed effects, and share the same spatial effect.
- Type IV: different parameters for fixed effects, and share the same spatial effect.
- Type V: share the same parameters for fixed effects, and the spatial effects between linear predictors are proportional.
- Type VI: different parameters for fixed effects, and the spatial effects between linear predictors are proportional.
- Type VII: share the same parameters for fixed effects, and different realizations of the spatial effect for each linear predictor.
- Type VIII: different parameters for fixed effects, and different realizations of the spatial effect for each linear predictor.

Note that in Type V and Type VI structures, realizations of the spatial field are the same, but a proportionality hyperparameter is added in two of the three linear predictors, denoted as $\alpha^{(1)}$ and $\alpha^{(2)}$. On the other hand, in Type VII and Type VIII structures, although realizations of random effects are different, they share the same hyperparameters. Table 7 provides an overview of the distinct structures, their associated latent fields, and the accompanying set of hyperparameters.

Table 8
Arabidopsis thaliana data: model selection criteria (DIC) and predictive performance measures (Logarithmic Score) with automatic groups construction for $m = 3, 5$, and 10.

Type	DIC	Logarithmic Score		
		$m = 3$	$m = 5$	$m = 10$
I	3353.403	1.894	1.897	1.904
II	3294.855	1.869	1.873	1.883
III	3202.082	1.784	1.780	1.795
IV	3145.198	1.757	1.756	1.778
V	3059.359	1.470	1.579	1.634
VI	3002.910	1.396	1.526	1.588
VII	2754.844	1.427	**1.513**	**1.567**
VIII	**2741.440**	**1.415**	**1.513**	**1.568**

4.2.3. LGOCV in compositional data

As highlighted in Martínez-Minaya and Rue (2024), excluding a category from a CoDa point during cross-validation procedures may not be practical due to the inherent constraint imposed by CoDa, where the sum of its components must equal one. This implies that the remaining categories hold valuable information about the category we intend to exclude. Consequently, the remaining log-ratio coordinates offer insights into the category we have removed during cross-validation. In this manner, the concept of 'friendship' emerges. Accordingly, we can assert that the first *alr*-coordinate of ith data point is associated with the second *alr*-coordinate of ith data point, thereby contributing pertinent information.

Hence, to perform cross-validation for the ith data point and the dth *alr*-coordinate, it becomes necessary to exclude the values associated with all *alr*-coordinates for that data point and for the corresponding group. To do this, the group $I_i^{(d)}$ must be constructed. If it is determined that the ith data point in its dth *alr*-coordinate is associated with the data point indexed as $i + 1$ in its dth *alr*-coordinate, the data point $i + 1$ will also be included in the set $I_i^{(d)}$ in all of its *alr*-coordinates. It should be noted that $I_i^{(d)}$ could be different from $I_i^{(d^*)}$, being $d \neq d^*$, as even though we are referring to the same compositional point i, we are modeling each *alr*-coordinate. So we need to compute LS measures for CoDa in the following way:

$$\text{LS} = -\frac{1}{N \cdot (D-1)} \sum_{d=1}^{D-1} \sum_{i=1}^{N} \log\left(\pi(\text{alr}(\boldsymbol{y})_i^{(d)} \mid \text{alr}(\boldsymbol{y})^{\bullet}_{-I_i^{(d)}})\right),$$

where $alr(\boldsymbol{y})_i^{(d)}$ is the observed vector for the ith data point and the dth *alr*-coordinate, and $alr(\boldsymbol{y})^{\bullet}_{-i}$ represents the observed data in *alr*-coordinates excluding the $I_i^{(d)}$ data points with its corresponding $D - 1$ *alr*-coordinates.

4.2.4. Results

Table 8 presents a comparison of various models used for estimating the membership proportions to the different genetic clusters of *Arabidopsis thaliana*. As in the previous section, the evaluation is based on model selection criteria and predictive performance measures. To compute comparable LS measures from the posterior predictive density functions, we use the groups derived from Type VIII model. We use values of $m = 3, 5$ and 10 for constructing the groups and compute performance measures based on LGOCV. Although, Type VII model seems to have a good predictive performance, DIC/WAIC and predictive measures based on LGOCV approaches consistently suggest that Type VIII model is the most suitable choice.

In Fig. 8, we depict automatic groups formed from the posterior correlations in the Type VIII model. Specifically, this representation pertains to the three *alr*-coordinates associated with the 88-th data point when $m = 10$. By employing the 'friendship' feature, we made the assumption that these three observations share a bond, meaning that if we intend to create groups for the first *alr*-coordinate of observation 88, we must not only include that *alr*-coordinate of observation $i = 88$ in $I_{88}^{(1)}$, but also the second and third *alr*-coordinates for that data point. As shown in Fig. 8, it becomes evident that the groups established for data point $i = 88$ differ for each *alr*-coordinate, i.e., $I_{88}^{(1)} \neq I_{88}^{(2)} \neq I_{88}^{(3)}$. This observation underscores not only our assumption of distinct realizations of the spatial effect for each *alr*-coordinate but also varying effects of the climatic variables. This is due to the fact that posterior correlations consider all components of the linear predictor.

4.3. Space–time models for the United Kingdom wind speed data

In what follows, we describe how to use the procedure outlined in Section 2 to perform cross-validation on space–time models for wind speed data in the United Kingdom (UK). We consider a dataset from a regularly updated 20th and 21st-century database from the European Climate Assessment & Dataset (ECA&D) project (Klein Tank et al., 2002). We select daily wind speed measurements, in meters per second (m/s), from July 2021 at 189 stations in Ireland and the UK. The locations of the stations are presented as colored dots in Fig. 9. The observations at each location are displayed with different colors in this figure as grouped time series over July 2021. The grouped time series are constructed by computing the average daily wind speed over the stations located within each spatial grid (outlined by dashed gray lines).

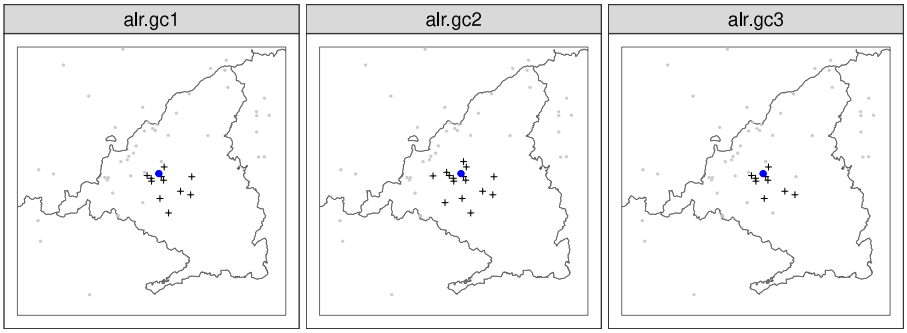


Fig. 8. Automatically constructed groups (with $m = 10$) under Type VIII model for data point $i = 88$ located in the region of Madrid. The blue circle represents the focal point, while the crosses indicate the points that belong to the constructed group $I_{88}^{(d)}$, $d = 1, \ldots, 3$.

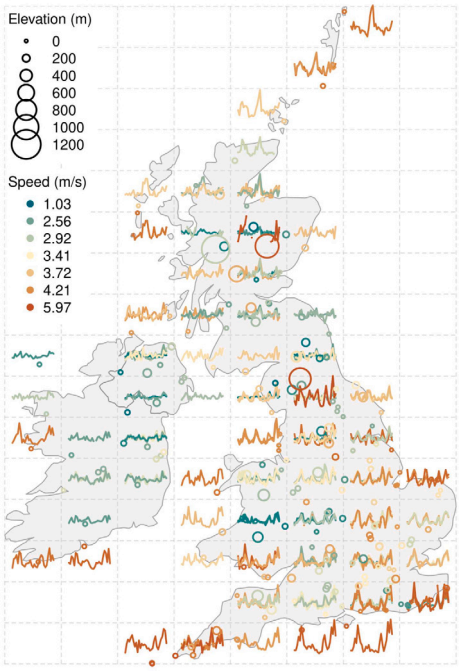


Fig. 9. Map displaying the station locations with the data presented as grouped time series (respect to the spatial grid overlayed in dashed gray lines) over the period of July 2021.

4.3.1. The models

We compare four models with different characterizations of the space–time dynamics in the UK wind speed data. Consider the process $Y(\mathbf{s}, t)$, which represents the wind speed at location $\mathbf{s}$ and at time t. We model observations $y(\mathbf{s}_i, t_i)$ from this process collected at site $\mathbf{s}_i$ and time t_i with a Gaussian distribution $\mathrm{N}(\mu(\mathbf{s}_i, t_i), \sigma_e^2)$, where σ_e^2 is a common variance parameter and the mean function is given by

$$\mu(\mathbf{s}, t) = \beta_0 + \beta_1 x(\mathbf{s}) + r(\mathbf{s}) + u(\mathbf{s}, t), \tag{4}$$

where β_0 is the intercept, which is constant in space and time, x is the elevation (in meters), and β_1 represents the associated regression coefficient. The $r(.)$ term is a spatial field, which accommodates regional persistent spatial variations and is modeled with a Matérn covariance using the stochastic partial differential equation (SPDE) approach described in Lindgren et al. (2011) with smoothness parameter fixed to $\alpha = 2$, range ρ_r, and marginal standard error σ_r, which need to be estimated. Space–time interactions are captured through the space–time random field $u(.)$, defined using the models proposed in Lindgren et al. (2023).

Table 9
Wind speed data: model selection criteria (DIC) and predictive performance measures (Logarithmic Score) with automatic groups construction for $m = 3$, $m = 5$, and $m = 10$.

Model for u	DIC	Logarithmic Score		
		m = 3	m = 5	m = 10
M_A	3499.78	6202.79	6692.26	7461.16
M_B	**1622.16**	**6054.93**	**6563.31**	**7413.71**
M_C	3433.40	6214.94	6678.89	7435.29
M_D	2078.38	7123.03	7587.39	8342.42

These models are defined as solutions to a diffusion-based family of equations that induces physically realistic processes, such as wind. Specifically, it is the solution to the following SPDE

$$\gamma_e(\gamma_s^2 - \Delta)^{\alpha_e/2}\left(\gamma_t\frac{\partial}{\partial t} + (\gamma_s^2 - \Delta)^{\alpha_s}\right)^{\alpha_t} u(s,t) = \mathcal{W}(s,t). \tag{5}$$

The power parameters α_t, α_s and α_e in Eq. (5) specify the smoothness of $u(.)$, whereas γ_s, γ_t, and γ_e govern the marginal properties and are related to dynamics of the process. We consider here four cases of the models proposed in Lindgren et al. (2023): Model M_A (with $\alpha_t = 1$, $\alpha_s = 0$, and $\alpha_e = 2$); Model M_B (with $\alpha_t = 1$, $\alpha_s = 2$, and $\alpha_e = 1$); Model M_C (with $\alpha_t = 2$, $\alpha_s = 0$, and $\alpha_e = 2$); and Model M_D (with $\alpha_t = 2$, $\alpha_s = 2$, and $\alpha_e = 0$). When $\alpha_s = 0$ (Models M_A and M_C), the space–time SPDE in Eq. (5) generates a precision matrix that can be factorized into two precision matrices, one for the purely spatial model and another for the purely temporal model. While separability is broadly used due to a substantial reduction of the computation time when working with large data sets with complex dependencies, it is usually too simplistic to model natural phenomena. Due to recent developments, with the SPDE approach both, separable or non-separable, have similar computation time.

For fitting the models, we considered a flat prior for β_0 and a Gaussian with zero mean and variance 0.01 for β_1. The penalized complexity prior proposed in Fuglstad et al. (2018), are used for the spatial and temporal range parameters of r and u, as well as for all the variance parameters σ_e^2, σ_r^2 (marginal variance of r), and σ_u^2 (marginal variance of u).

4.3.2. Results

We use the methods described in Section 2 to perform LGOCV on the four models presented above. For all the models, we set the groups to those generated from model M_A. Fig. 10 illustrates the groups for $m = 10$, at time 1 for seven locations from different parts of the spatial domain. Even though m is fixed to 10, the total number of observations in each group is always larger than 10 (see the last row of the summary table on the top left of this figure). This is due to several observation pairs having very similar correlation values. For instance, the group for station 'id' = 5 is of size 17 and for 'id' = 131 it is of size 50, which is the number we set as the maximum group size. In the top left table, for each station, the removed neighbors are classified into *self* (same location and time), *time* (same location but different times), *spatial* (same time but different locations), or *space–time* (different location and different time). Stations located in isolated places such as 'id' = 131 tend to select more neighbors in time. Indeed, only three spatial neighbors were selected and the entire time series of this station was left out, that is fifteen observations before and fifteen after the selected time. On the other hand, stations that have several neighbors close by such as 'id' = 18 tend to mostly include spatial and space–time neighbors.

Finally, Table 9 shows the DIC for the different models and the LS values for $m = 3$, $m = 5$, and $m = 10$. Model M_B provides the best fit with the lowest DIC. This model also performed the best for prediction purposes, as shown by the LS values for all level sets. Model M_D has the second lowest DIC but the worst LS, indicating that it is too smooth to predict the highly variable wind trajectories (see Fig. 9).

5. Discussion

Cross-validation techniques are at the core of any prediction task as they reveal the adequacy of a statistical model to unobserved data. Among the several cross-validation methods developed over the years, LOOCV stands out as the prevailing choice in numerous practical applications due to its simplicity and straightforward implementation. Although widely used, this method is only optimal under the assumption of independent and identically distributed (iid) data, an assumption rarely verified in many practical applications. In contrast, the newly developed LGOCV method handles non iid data with multiple correlation structures by automatically excluding a determined set of data points from the training set, expanding beyond the practice of solely excluding individual observations. The R-INLA package provides a convenient implementation of LGOCV for latent Gaussian models that is highly efficient and readily accessible.

In this paper, we showed that the automatic group construction procedure for LGOCV provides a robust measure of predictive performance in many settings where dependencies in space and/or time are present. Our simulation studies and real data analysis underscore that a common practice in modeling dependent data often leads to erroneous model selection and interpretations. In contrast, the LGOCV method, integrated into the R-INLA package, stands as an improved alternative to LOOCV. In the model selection process, the LGOCV framework should be the primary consideration, especially when evaluating the predictive extrapolation performance of models.

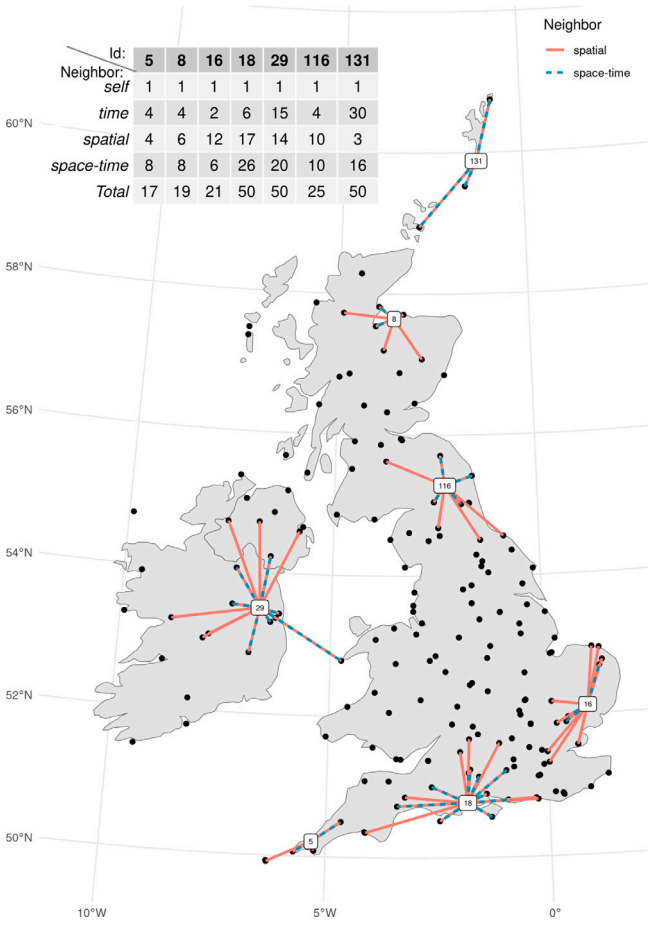


Fig. 10. Location of the stations and examples of seven observations at time 16 with their respective automatic selected groups from the LGOCV method.

Acknowledgments

Open access funding provided by Universidad Pública de Navarra. This research has been supported by project PID2020-113125RB-I00/MCIN/AEI/10.13039/501100011033 for Adin, A., and by project PID2020-115882RB-I00 for Martínez-Minaya, J. We would like to thank the valuable comments made by two anonymous reviewers that have contributed to clarify some aspects of this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spasta.2024.100843.

References

Adin, A., Goicoa, T., Hodges, J.S., Schnell, P., Ugarte, M.D., 2023a. Alleviating confounding in spatio-temporal areal models with an application on crimes against women in India. Stat. Model. 23 (1), 9–30. http://dx.doi.org/10.1177/1471082X211015452.

Adin, A., Goicoa, Ugarte, M.D., 2023b. Multivariate disease mapping models to uncover hidden relationships between different cancer sites. In: Larriba, Y. (Ed.), Statistical Methods At the Forefront of Biomedical Advances. Springer International Publishing, pp. 1–20. http://dx.doi.org/10.1007/978-3-031-32729-2_1.

Aguilera, A., Bautista, F., Gutiérrez-Ruiz, M., Ceniceros-Gómez, A.E., Cejudo, R., Goguitchaichvili, A., 2021. Heavy metal pollution of street dust in the largest city of Mexico, sources and health risk assessment. Environ. Monit. Assess. 193 (4), 1–16. http://dx.doi.org/10.1007/s10661-021-09344-z.

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4, 40–79. http://dx.doi.org/10.1214/09-SS054.

Bergmeir, C., Benítez, J.M., 2012. On the use of cross-validation for time series predictor evaluation. Inform. Sci. 191, 192–213. http://dx.doi.org/10.1016/j.ins.2011.12.028.

Buccianti, A., Grunsky, E., 2014. Compositional data analysis in geochemistry: Are we sure to see what really occurs during natural processes? J. Geochem. Explor. 141, 1–5. http://dx.doi.org/10.1016/j.gexplo.2014.03.022.

Bürkner, P.-C., Gabry, J., Vehtari, A., 2021. Efficient leave-one-out cross-validation for Bayesian non-factorized normal and student-t models. Comput. Statist. 36 (2), 1243–1261. http://dx.doi.org/10.1007/s00180-020-01045-4.

Creus Martí, I., Moya, A., Santonja, F., 2022. Bayesian hierarchical compositional models for analysing longitudinal abundance data from microbiome studies. Complexity 2022, http://dx.doi.org/10.1155/2022/4907527.

Douma, J.C., Weedon, J.T., 2019. Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. Methods Ecol. Evol. 10 (9), 1412–1430. http://dx.doi.org/10.1111/2041-210X.13234.

Dumuid, D., Stanford, T.E., Martin-Fernández, J.-A., Pedišić, Ž., Maher, C.A., Lewis, L.K., Hron, K., Katzmarzyk, P.T., Chaput, J.-P., Fogelholm, M., et al., 2018. Compositional data analysis for physical activity, sedentary time and sleep research. Stat. Methods Med. Res. 27 (12), 3726–3738. http://dx.doi.org/10.1177/09622802177108.

Engle, M.A., Rowan, E.L., 2014. Geochemical evolution of produced waters from hydraulic fracturing of the marcellus shale, northern appalachian basin: A multivariate compositional data analysis approach. Int. J. Coal Geol. 126, 45–56. http://dx.doi.org/10.1016/j.coal.2013.11.010.

Etxeberria, J., Goicoa, T., Ugarte, M.D., 2023. Using mortality to predict incidence for rare and lethal cancers in very small areas. Biom. J. 65 (3), 2200017. http://dx.doi.org/10.1002/bimj.202200017.

Fairclough, S.J., Dumuid, D., Mackintosh, K.A., Stone, G., Dagger, R., Stratton, G., Davies, I., Boddy, L.M., 2018. Adiposity, fitness, health-related quality of life and the reallocation of time between children's school day activity behaviours: A compositional data analysis. Prevent. Med. Rep. 11, 254–261. http://dx.doi.org/10.1016/j.pmedr.2018.07.011.

Fuglstad, G.A., Simpson, D., Lindgren, F., Rue, H., 2018. Constructing priors that penalize the complexity of Gaussian random fields. J. Amer. Statist. Assoc. 114 (525), 445–452. http://dx.doi.org/10.1080/01621459.2017.1415907.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis. Chapman and Hall/CRC.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 102 (477), 359–378. http://dx.doi.org/10.1198/016214506000001437.

Goicoa, T., Adin, A., Ugarte, M.D., Hodges, J.S., 2018. In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. Stoch. Environ. Res. Risk Assess. 32 (3), 749–770. http://dx.doi.org/10.1007/s00477-017-1405-0.

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

Held, L., Natário, I., Fenton, S.E., Rue, H., Becker, N., 2005. Towards joint disease mapping. Stat. Methods Med. Res. 14 (1), 61–82. http://dx.doi.org/10.1191/0962280205sm389oa.

Held, L., Schrodle, B., Rue, H., 2010. Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. In: Statistical Modelling and Regression Structures. Springer, pp. 111–131.

Klein Tank, A., Wijngaard, J., Können, G., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., et al., 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the European climate assessment. Int. J. Climatol.: J. R. Meteorol. Soc. 22 (12), 1441–1453. http://dx.doi.org/10.1002/joc.773.

Knorr-Held, L., 2000. Bayesian modelling of inseparable space-time variation in disease risk. Stat. Med. 19 (17–18), 2555–2567. http://dx.doi.org/10.1002/1097-0258(20000915/30)19:17/18<2555::AID-SIM587>3.0.CO;2-#.

Knorr-Held, L., Best, N.G., 2001. A shared component model for detecting joint and selective clustering of two diseases. J. R. Stat. Soc. Ser. A: Stat. Soc. 164 (1), 73–85. http://dx.doi.org/10.1111/1467-985X.00187.

Kobal, M., Kastelec, D., Eler, K., 2017. Temporal changes of forest species composition studied by compositional data approach. iForest-Biogeosci. Forest. 10 (4), 729–738. http://dx.doi.org/10.3832/ifor2187-010.

Kuhn, M., Johnson, K., et al., 2013. Applied Predictive Modeling, vol. 26, Springer.

Lindgren, F., Bakka, H., Bolin, D., Krainski, E., Rue, H., 2023. A diffusion-based spatio-temporal extension of Gaussian Matérn fields. http://dx.doi.org/10.48550/arXiv.2006.04917.

Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (4), 423–498. http://dx.doi.org/10.1111/j.1467-9868.2011.00777.x.

Liu, Z., Rue, H., 2022. Leave-group-out cross-validation for latent Gaussian models. http://dx.doi.org/10.48550/arXiv.2210.04482.

MacNab, Y.C., 2018. Some recent work on multivariate Gaussian Markov random fields. Test 27 (3), 497–541. http://dx.doi.org/10.1007/s11749-018-0605-3.

Martínez-Minaya, J., Conesa, D., Fortin, M.-J., Alonso-Blanco, C., Picó, F.X., Marcer, A., 2019. A hierarchical Bayesian beta regression approach to study the effects of geographical genetic structure and spatial autocorrelation on species distribution range shifts. Mol. Ecol. Resour. 19 (4), 929–943. http://dx.doi.org/10.1111/1755-0998.13024.

Martínez-Minaya, J., Lindgren, F., López-Quílez, A., Simpson, D., Conesa, D., 2023. The integrated nested Laplace approximation for fitting Dirichlet regression models. J. Comput. Graph. Statist. 32 (3), 805–823. http://dx.doi.org/10.1080/10618600.2022.2144330.

Martínez-Minaya, J., Rue, H., 2024. A flexible Bayesian tool for CoDa mixed models: Logistic-normal distribution with Dirichlet covariance. Stat. Comput. 34 (3), 116. http://dx.doi.org/10.1007/s11222-024-10427-3.

Mota-Bertran, A., Saez, M., Coenders, G., 2022. Compositional and Bayesian inference analysis of the concentrations of air pollutants in Catalonia, Spain. Environ. Res. 204, 112388. http://dx.doi.org/10.1016/j.envres.2021.112388.

Rabinowicz, A., Rosset, S., 2022. Cross-validation for correlated data. J. Amer. Statist. Assoc. 117 (538), 718–731. http://dx.doi.org/10.1080/01621459.2020.1801451.

Riebler, A., Sørbye, S.H., Simpson, D., Rue, H., 2016. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. Stat. Methods Med. Res. 25 (4), 1145–1165. http://dx.doi.org/10.1177/0962280216660421.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., et al., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40 (8), 913–929. http://dx.doi.org/10.1111/ecog.02881.

Rue, H., Martino, S., Chopin, N., 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (2), 319–392. http://dx.doi.org/10.1111/j.1467-9868.2008.00700.x.

Shi, P., Zhang, A., Li, H., et al., 2016. Regression analysis for microbiome compositional data. Ann. Appl. Stat. 10 (2), 1019–1040. http://dx.doi.org/10.1214/16-AOAS928.

Simpson, D., Rue, H., Riebler, A., Martins, T.G., Sørbye, S.H., 2017. Penalising model component complexity: A principled, practical approach to constructing priors. Statist. Sci. 32 (1), 1–28. http://dx.doi.org/10.1214/16-STS576.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (4), 583–639. http://dx.doi.org/10.1111/1467-9868.00353.

Tsilimigras, M.C., Fodor, A.A., 2016. Compositional data analysis of the microbiome: Fundamentals, tools, and challenges. Ann. Epidemiol. 26 (5), 330–335. http://dx.doi.org/10.1016/j.annepidem.2016.03.002.

Ugarte, M.D., Adin, A., Goicoa, T., 2017. One-dimensional, two-dimensional, and three dimensional B-splines to specify space–time interactions in Bayesian disease mapping: Model fitting and model identifiability. Spat. Stat. 22, 451–468. http://dx.doi.org/10.1016/j.spasta.2017.04.002.

Van Niekerk, J., Krainski, E., Rustand, D., Rue, H., 2023. A new avenue for Bayesian inference with INLA. Comput. Statist. Data Anal. 181, 107692. http://dx.doi.org/10.1016/j.csda.2023.107692.

Vicente, G., Goicoa, T., Fernandez-Rasines, P., Ugarte, M.D., 2020a. Crime against women in India: Unveiling spatial patterns and temporal trends of dowry deaths in the districts of Uttar Pradesh. J. R. Stat. Soc. Ser. A: Stat. Soc. 183 (2), 655–679. http://dx.doi.org/10.1111/rssa.12545.

Vicente, G., Goicoa, T., Ugarte, M.D., 2020b. Bayesian inference in multivariate spatio-temporal areal models using INLA: Analysis of gender-based violence in small areas. Stoch. Environ. Res. Risk Assess. 34, 1421–1440. http://dx.doi.org/10.1007/s00477-020-01808-x.

Washburne, A.D., Silverman, J.D., Leff, J.W., Bennett, D.J., Darcy, J.L., Mukherjee, S., Fierer, N., David, L.A., 2017. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. PeerJ 5, e2969. http://dx.doi.org/10.7717/peerj.2969.

Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. 11 (116), 3571–3594, URL http://jmlr.org/papers/v11/watanabe10a.html.