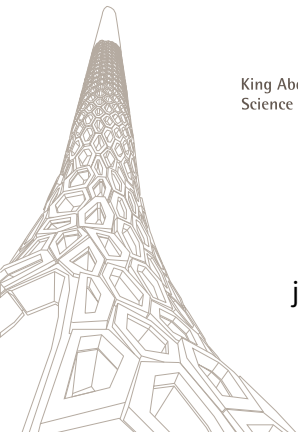


Cross-validation for hierarchical models



King Abdullah University of
Science and Technology



جامعة الملك عبد الله
للعلوم والتقنية

Janet van Niekerk
janet.vanNiekerk@kaust.edu.sa
July 2025





- 1 LOOCV
- 2 Problem statement
- 3 LGOCV
- 4 Details
- 5 Examples
 - Time series models
 - Spatial models



Introduction to CV

Suppose we have data,

$$\mathbf{y} = \{y_i\} \sim \pi_T(\mathbf{y})$$

Our objective is to determine how well a fitted model can predict a new observation, \tilde{y} from this true distribution.

In the Bayesian context, we use the posterior predictive distribution $\pi(y|\mathbf{y})$ to predict \tilde{y} sampled from $\pi_T(\mathbf{y})$.

Using the logarithmic score, we can compute $E_{\tilde{y}}[\log \pi(\tilde{y}|\mathbf{y})]$ as a metric for prediction ability.



Introduction to CV

Suppose we have data,

$$\mathbf{y} = \{y_i\} \sim \pi_T(\mathbf{y})$$

Our objective is to determine how well a fitted model can predict a new observation, \tilde{y} from this true distribution.

In the Bayesian context, we use the posterior predictive distribution $\pi(y|\mathbf{y})$ to predict \tilde{y} sampled from $\pi_T(\mathbf{y})$.

Using the logarithmic score, we can compute $E_{\tilde{y}}[\log \pi(\tilde{y}|\mathbf{y})]$ as a metric for prediction ability.



Introduction to CV

Suppose we have data,

$$\mathbf{y} = \{y_i\} \sim \pi_T(\mathbf{y})$$

Our objective is to determine how well a fitted model can predict a new observation, \tilde{y} from this true distribution.

In the Bayesian context, we use the posterior predictive distribution $\pi(y|\mathbf{y})$ to predict \tilde{y} sampled from $\pi_T(\mathbf{y})$.

Using the logarithmic score, we can compute $E_{\tilde{y}}[\log \pi(\tilde{y}|\mathbf{y})]$ as a metric for prediction ability.



The informal interpretation of LOOCV is that it mimics “using \mathbf{y} to predict $\tilde{\mathbf{y}}$ ” by “using \mathbf{y}_{-i} to predict y_i ”.

This intuitive interpretation is then used to justify, often implicitly, the use of LOOCV as a “default” way to evaluate predictive performance.



The informal interpretation of LOOCV is that it mimics “using \mathbf{y} to predict $\tilde{\mathbf{y}}$ ” by “using \mathbf{y}_{-i} to predict y_i ”.

This intuitive interpretation is then used to justify, often implicitly, the use of LOOCV as a “default” way to evaluate predictive performance.



Time series models

Assume data $\mathbf{y} = \{y_1, y_2, \dots, y_T\}$ is a time-series, observed sequentially at time $1, 2, \dots, T$. The inherent prediction task is to predict future values, given the temporal nature of the data. We can predict a new observation at $k \geq 1$ steps into the future by $\pi(y_{T+k}|y_1, \dots, y_T)$.

In this example, the LOOCV will be computed from

$$\pi(y_t|y_1, \dots, y_{t-1}, y_{t+1}, \dots, y_T), \quad t = 1, \dots, T,$$

Leave-future-out cross-validation (LFOCV):

$$\sum_{T'=T_0}^{T-k} \log \pi(y_{T'+k}|y_1, \dots, y_{T'}),$$

where T' starts from time $T_0 > 1$ as we need some data to estimate the model.



Time series models

Assume data $\mathbf{y} = \{y_1, y_2, \dots, y_T\}$ is a time-series, observed sequentially at time $1, 2, \dots, T$. The inherent prediction task is to predict future values, given the temporal nature of the data. We can predict a new observation at $k \geq 1$ steps into the future by $\pi(y_{T+k}|y_1, \dots, y_T)$.

In this example, the LOOCV will be computed from

$$\pi(y_t|y_1, \dots, y_{t-1}, y_{t+1}, \dots, y_T), \quad t = 1, \dots, T,$$

Leave-future-out cross-validation (LFOCV):

$$\sum_{T'=T_0}^{T-k} \log \pi(y_{T'+k}|y_1, \dots, y_{T'}),$$

where T' starts from time $T_0 > 1$ as we need some data to estimate the model.



Time series models

Assume data $\mathbf{y} = \{y_1, y_2, \dots, y_T\}$ is a time-series, observed sequentially at time $1, 2, \dots, T$. The inherent prediction task is to predict future values, given the temporal nature of the data. We can predict a new observation at $k \geq 1$ steps into the future by $\pi(y_{T+k}|y_1, \dots, y_T)$.

In this example, the LOOCV will be computed from

$$\pi(y_t|y_1, \dots, y_{t-1}, y_{t+1}, \dots, y_T), \quad t = 1, \dots, T,$$

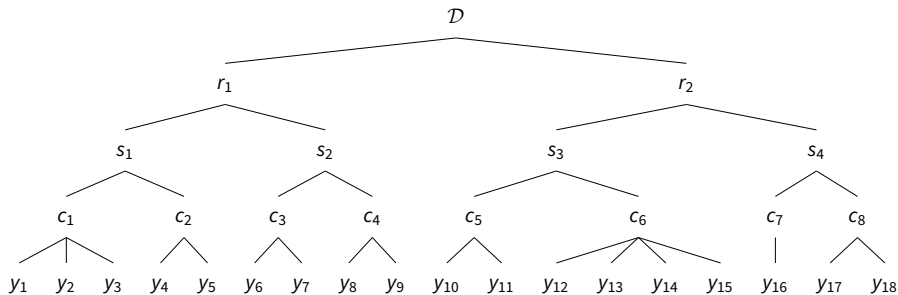
Leave-future-out cross-validation (LFOCV):

$$\sum_{T'=T_0}^{T-k} \log \pi(y_{T'+k}|y_1, \dots, y_{T'}),$$

where T' starts from time $T_0 > 1$ as we need some data to estimate the model.



Multilevel models

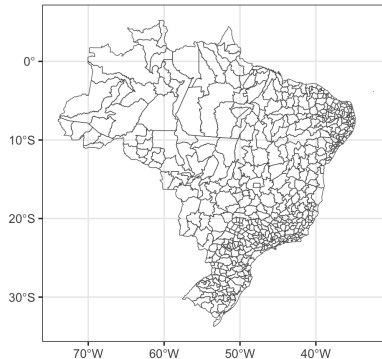




Spatial models

$$y_i | \eta_i \sim \text{Poisson}(E_i \exp(\eta_i))$$

$$\eta_i = \mu + u_i, \quad u_i \text{ a smooth process in space}$$





What to do?

Non-exchangeable data → the prediction task implicitly defined through LOOCV may be less appropriate, as it leans more towards assessing imputing qualities and short range predictions

What we want from CV:

- Evaluate predictive performance for longer range as is usually implied by "out-of-sample" prediction
- Model-specific hold-out set
- Efficient computation of a score



What to do?

Non-exchangeable data → the prediction task implicitly defined through LOOCV may be less appropriate, as it leans more towards assessing imputing qualities and short range predictions

What we want from CV:

- Evaluate predictive performance for longer range as is usually implied by "out-of-sample" prediction
- Model-specific hold-out set
- Efficient computation of a score



What to do?

Non-exchangeable data → the prediction task implicitly defined through LOOCV may be less appropriate, as it leans more towards assessing imputing qualities and short range predictions

What we want from CV:

- Evaluate predictive performance for longer range as is usually implied by "out-of-sample" prediction
- Model-specific hold-out set
- Efficient computation of a score



What to do?

Non-exchangeable data → the prediction task implicitly defined through LOOCV may be less appropriate, as it leans more towards assessing imputing qualities and short range predictions

What we want from CV:

- Evaluate predictive performance for longer range as is usually implied by "out-of-sample" prediction
- Model-specific hold-out set
- Efficient computation of a score



What to do?

Non-exchangeable data → the prediction task implicitly defined through LOOCV may be less appropriate, as it leans more towards assessing imputing qualities and short range predictions

What we want from CV:

- Evaluate predictive performance for longer range as is usually implied by "out-of-sample" prediction
- Model-specific hold-out set
- Efficient computation of a score



LGOCV

Leave-group-out cross-validation (LGOCV):

$$u_{\text{LGOCV}} = \frac{1}{n} \sum_{i=1}^n \log(\pi(y_i | \mathbf{y}_{-l_i})). \quad (1)$$

Here, the *group* (denoted by l_i) is an index set including i . This configuration facilitates that the pair (y_i, \mathbf{y}_{-l_i}) mimics a specified prediction task. In the multilevel model, predicting a student's grade from an unseen class necessitates that l_i includes i and all observations from student i 's class. However, more complex models, such as models containing both time series and hierarchical elements, pose challenges when defining a natural prediction task. We need some structure!



LGOCV

Leave-group-out cross-validation (LGOCV):

$$u_{\text{LGOCV}} = \frac{1}{n} \sum_{i=1}^n \log(\pi(y_i | \mathbf{y}_{-l_i})). \quad (1)$$

Here, the *group* (denoted by l_i) is an index set including i . This configuration facilitates that the pair (y_i, \mathbf{y}_{-l_i}) mimics a specified prediction task. In the multilevel model, predicting a student's grade from an unseen class necessitates that l_i includes i and all observations from student i 's class. However, more complex models, such as models containing both time series and hierarchical elements, pose challenges when defining a natural prediction task. We need some structure!



LGOCV

Leave-group-out cross-validation (LGOCV):

$$u_{\text{LGOCV}} = \frac{1}{n} \sum_{i=1}^n \log(\pi(y_i | \mathbf{y}_{-l_i})). \quad (1)$$

Here, the *group* (denoted by l_i) is an index set including i . This configuration facilitates that the pair (y_i, \mathbf{y}_{-l_i}) mimics a specified prediction task. In the multilevel model, predicting a student's grade from an unseen class necessitates that l_i includes i and all observations from student i 's class. However, more complex models, such as models containing both time series and hierarchical elements, pose challenges when defining a natural prediction task. We need some structure!



LGOCV

Leave-group-out cross-validation (LGOCV):

$$u_{\text{LGOCV}} = \frac{1}{n} \sum_{i=1}^n \log(\pi(y_i | \mathbf{y}_{-l_i})). \quad (1)$$

Here, the *group* (denoted by l_i) is an index set including i . This configuration facilitates that the pair (y_i, \mathbf{y}_{-l_i}) mimics a specified prediction task. In the multilevel model, predicting a student's grade from an unseen class necessitates that l_i includes i and all observations from student i 's class.

However, more complex models, such as models containing both time series and hierarchical elements, pose challenges when defining a natural prediction task. We need some structure!



LGOCV

Leave-group-out cross-validation (LGOCV):

$$u_{\text{LGOCV}} = \frac{1}{n} \sum_{i=1}^n \log(\pi(y_i | \mathbf{y}_{-l_i})). \quad (1)$$

Here, the *group* (denoted by l_i) is an index set including i . This configuration facilitates that the pair (y_i, \mathbf{y}_{-l_i}) mimics a specified prediction task. In the multilevel model, predicting a student's grade from an unseen class necessitates that l_i includes i and all observations from student i 's class. However, more complex models, such as models containing both time series and hierarchical elements, pose challenges when defining a natural prediction task. We need some structure!



LGOCV

Leave-group-out cross-validation (LGOCV):

$$u_{\text{LGOCV}} = \frac{1}{n} \sum_{i=1}^n \log(\pi(y_i | \mathbf{y}_{-l_i})). \quad (1)$$

Here, the *group* (denoted by l_i) is an index set including i . This configuration facilitates that the pair (y_i, \mathbf{y}_{-l_i}) mimics a specified prediction task. In the multilevel model, predicting a student's grade from an unseen class necessitates that l_i includes i and all observations from student i 's class. However, more complex models, such as models containing both time series and hierarchical elements, pose challenges when defining a natural prediction task. We need some structure!



LGM

Data \mathbf{y} and covariates \mathbf{A} ,

$$\begin{aligned} y_i | \eta_i, \boldsymbol{\theta} &\sim \pi(y_i | \eta_i, \boldsymbol{\theta}), \\ \eta &= \mathbf{A}\mathbf{f}, \quad \mathbf{f} | \boldsymbol{\theta} \sim N(0, \mathbf{P}_{\mathbf{f}}(\boldsymbol{\theta})), \quad \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}). \end{aligned} \tag{2}$$

The model is quite general because \mathbf{f} can combine many modeling components, including linear model, spatial components, temporal components, spline components, etc. It is also common with linear constraints on the latent effects \mathbf{f} .

Efficient full Bayesian inference using INLA



LGM

Data \mathbf{y} and covariates \mathbf{A} ,

$$\begin{aligned} y_i | \eta_i, \boldsymbol{\theta} &\sim \pi(y_i | \eta_i, \boldsymbol{\theta}), \\ \eta &= \mathbf{A}\mathbf{f}, \quad \mathbf{f} | \boldsymbol{\theta} \sim N(0, \mathbf{P}_{\mathbf{f}}(\boldsymbol{\theta})), \quad \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}). \end{aligned} \tag{2}$$

The model is quite general because \mathbf{f} can combine many modeling components, including linear model, spatial components, temporal components, spline components, etc. It is also common with linear constraints on the latent effects \mathbf{f} .

Efficient full Bayesian inference using INLA



Data \mathbf{y} and covariates \mathbf{A} ,

$$\begin{aligned} y_i | \eta_i, \boldsymbol{\theta} &\sim \pi(y_i | \eta_i, \boldsymbol{\theta}), \\ \boldsymbol{\eta} &= \mathbf{A}\mathbf{f}, \quad \mathbf{f} | \boldsymbol{\theta} \sim N(0, \mathbf{P}_{\mathbf{f}}(\boldsymbol{\theta})), \quad \boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}). \end{aligned} \tag{2}$$

The model is quite general because \mathbf{f} can combine many modeling components, including linear model, spatial components, temporal components, spline components, etc. It is also common with linear constraints on the latent effects \mathbf{f} .

Efficient full Bayesian inference using INLA



Prediction task

How can we define a generic prediction task?

Out of sample prediction performance → mimic long range prediction ...

So which points to remove?



Prediction task

How can we define a generic prediction task?

Out of sample prediction performance → mimic long range prediction ...

So which points to remove?



Prediction task

How can we define a generic prediction task?

Out of sample prediction performance → mimic long range prediction ...

So which points to remove?

- Which y to leave out (and how many)?
- Computing $\pi(y_i | \mathbf{y}_{-l_i})$ efficiently.



Automatic group construction

In LGMs, the linear predictors, η , represent the underlining data generation process of data

$$\mathbf{R}_{\text{post}} = \text{corr}(\eta)$$

Sort the η_i according to \mathbf{R}_{post} to make the prediction "more and more" out of sample.

Bonus: model-based clusters



Automatic group construction

In LGMs, the linear predictors, η , represent the underlining data generation process of data

$$\mathbf{R}_{\text{post}} = \text{corr}(\eta)$$

Sort the η_i according to \mathbf{R}_{post} to make the prediction "more and more" out of sample.

Bonus: model-based clusters



Automatic group construction

In LGMs, the linear predictors, η , represent the underlining data generation process of data

$$\mathbf{R}_{\text{post}} = \text{corr}(\eta)$$

Sort the $\eta_{i,j}$ according to \mathbf{R}_{post} to make the prediction "more and more" out of sample.

Bonus: model-based clusters



Automatic group construction

In LGMs, the linear predictors, η , represent the underlining data generation process of data

$$\mathbf{R}_{\text{post}} = \text{corr}(\eta)$$

Sort the $\eta_{i,j}$ according to \mathbf{R}_{post} to make the prediction "more and more" out of sample.

Bonus: model-based clusters



Automatic group construction

Algorithm 1: Find groups for all data points

```

1 Input: A correlation matrix choice  $\mathbf{R}$  (posterior correlation is the default), Number of level sets  $m$ ;
2 Output: A list containing the groups for all data points;
3 Calculate  $\mathbf{R}$  from the model;
4  $N \leftarrow$  number of rows in  $\mathbf{R}$ ;
5 groups  $\leftarrow$  initialize  $N$  empty lists;
6 for  $i = 1$  to  $N$  do
7      $\mathbf{r} \leftarrow$  absolute values of the  $i$ -th row of  $\mathbf{R}$ ;
8     ordered indices  $\leftarrow$  indices of  $\mathbf{r}$  sorted by value in decreasing order;
9     current absolute correlation  $\leftarrow 1$ ;
10     $k \leftarrow 1$ ;
11    for  $j = 1$  to  $m$  do
12        while current absolute correlation  $\neq \mathbf{r}[\text{ordered indices}[k]]$  do
13            groups[ $i$ ].append(ordered indices[ $k$ ]);
14             $k \leftarrow k + 1$ ;
15        end
16        current absolute correlation  $\leftarrow \mathbf{r}[\text{ordered indices}[k]]$ ;
17    end
18 end
19 return groups;
  
```



Posterior predictive density computation

$$\pi(y_i|\mathbf{y}_{-l_i}) = \int_{\boldsymbol{\theta}} \pi(y_i|\boldsymbol{\theta}, \mathbf{y}_{-l_i})\pi(\boldsymbol{\theta}|\mathbf{y}_{-l_i})d\boldsymbol{\theta} \quad (3)$$

$$\pi(y_i|\boldsymbol{\theta}, \mathbf{y}_{-l_i}) = \int \pi(y_i|\eta_i, \boldsymbol{\theta})\pi(\eta_i|\boldsymbol{\theta}, \mathbf{y}_{-l_i})d\eta_i. \quad (4)$$



Posterior predictive density computation

$$\pi(\eta_i | \boldsymbol{\theta}, \mathbf{y}_{-l_i}) \approx \pi_G(\eta_i | \boldsymbol{\theta}, \mathbf{y}_{-l_i}) = N(\mu_{-l_i}, \boldsymbol{\Sigma}_{-l_i}).$$

We have

$$\pi_G(\eta | \boldsymbol{\theta}, \mathbf{y}) = N(\mu, \boldsymbol{\Sigma})$$

from INLA.

Now we use (inverse) Bayes' theorem, because : $\pi_G(\eta | \boldsymbol{\theta}, \mathbf{y})$ can be viewed as a posterior based on likelihood $\pi(\mathbf{y}_l | \eta_l)$ and prior $\pi_G(\eta_l | \boldsymbol{\theta}, \mathbf{y}_{-l_i})$.

Complications like constraints....



Posterior predictive density computation

$$\pi(\eta_i | \boldsymbol{\theta}, \mathbf{y}_{-l_i}) \approx \pi_G(\eta_i | \boldsymbol{\theta}, \mathbf{y}_{-l_i}) = N(\mu_{-l_i}, \boldsymbol{\Sigma}_{-l_i}).$$

We have

$$\pi_G(\eta | \boldsymbol{\theta}, \mathbf{y}) = N(\mu, \boldsymbol{\Sigma})$$

from INLA.

Now we use (inverse) Bayes' theorem, because : $\pi_G(\eta | \boldsymbol{\theta}, \mathbf{y})$ can be viewed as a posterior based on likelihood $\pi(\mathbf{y}_l | \eta_l)$ and prior $\pi_G(\eta_l | \boldsymbol{\theta}, \mathbf{y}_{-l_i})$.

Complications like constraints....



Posterior predictive density computation

$$\pi(\eta_i | \boldsymbol{\theta}, \mathbf{y}_{-l_i}) \approx \pi_G(\eta_i | \boldsymbol{\theta}, \mathbf{y}_{-l_i}) = N(\mu_{-l_i}, \boldsymbol{\Sigma}_{-l_i}).$$

We have

$$\pi_G(\eta | \boldsymbol{\theta}, \mathbf{y}) = N(\mu, \boldsymbol{\Sigma})$$

from INLA.

Now we use (inverse) Bayes' theorem, because : $\pi_G(\eta | \boldsymbol{\theta}, \mathbf{y})$ can be viewed as a posterior based on likelihood $\pi(\mathbf{y}_l | \eta_l)$ and prior $\pi_G(\eta_l | \boldsymbol{\theta}, \mathbf{y}_{-l_l})$.

Complications like constraints....



Posterior predictive density computation

$$\pi(\eta_i | \boldsymbol{\theta}, \mathbf{y}_{-l_i}) \approx \pi_G(\eta_i | \boldsymbol{\theta}, \mathbf{y}_{-l_i}) = N(\mu_{-l_i}, \boldsymbol{\Sigma}_{-l_i}).$$

We have

$$\pi_G(\eta | \boldsymbol{\theta}, \mathbf{y}) = N(\mu, \boldsymbol{\Sigma})$$

from INLA.

Now we use (inverse) Bayes' theorem, because : $\pi_G(\eta | \boldsymbol{\theta}, \mathbf{y})$ can be viewed as a posterior based on likelihood $\pi(\mathbf{y}_l | \eta_l)$ and prior $\pi_G(\eta_l | \boldsymbol{\theta}, \mathbf{y}_{-l_i})$.

Complications like constraints....



Posterior predictive density computation

$$\pi(\eta_i | \boldsymbol{\theta}, \mathbf{y}_{-l_i}) \approx \pi_G(\eta_i | \boldsymbol{\theta}, \mathbf{y}_{-l_i}) = N(\mu_{-l_i}, \boldsymbol{\Sigma}_{-l_i}).$$

We have

$$\pi_G(\eta | \boldsymbol{\theta}, \mathbf{y}) = N(\mu, \boldsymbol{\Sigma})$$

from INLA.

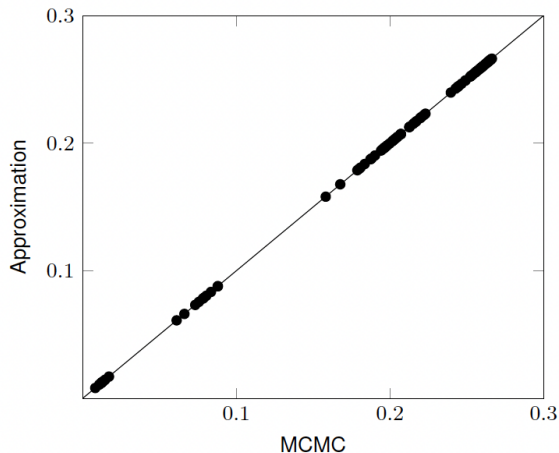
Now we use (inverse) Bayes' theorem, because : $\pi_G(\eta | \boldsymbol{\theta}, \mathbf{y})$ can be viewed as a posterior based on likelihood $\pi(\mathbf{y}_l | \eta_l)$ and prior $\pi_G(\eta_l | \boldsymbol{\theta}, \mathbf{y}_{-l_i})$.

Complications like constraints....



vs MCMC

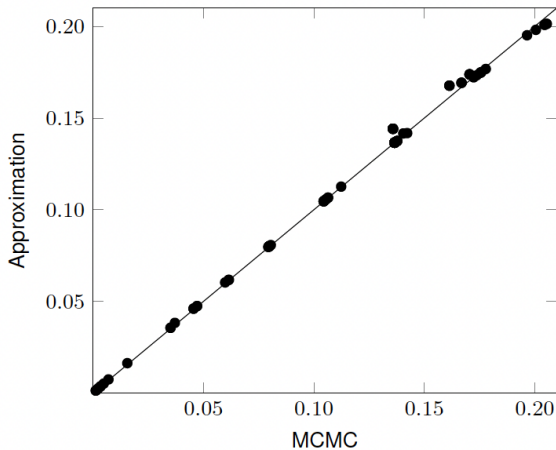
(b) Comparison for Gaussian response





vs MCMC

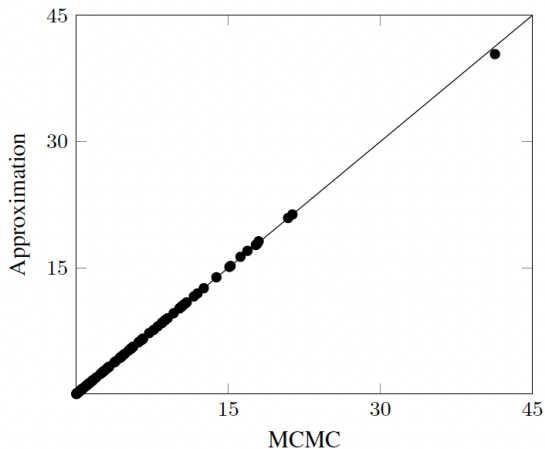
(d) Comparison for binomial response





vs MCMC

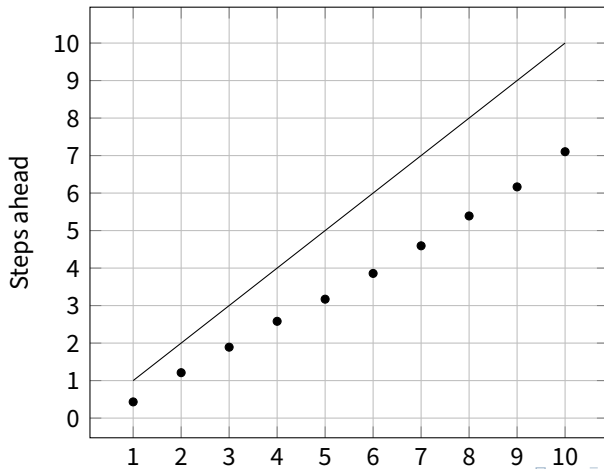
(f) Comparison for exponential response





Time series models

(b) Correspondence between LGOCV and LFOCV





Cancer risk in Germany

We fit the following model on the data set:

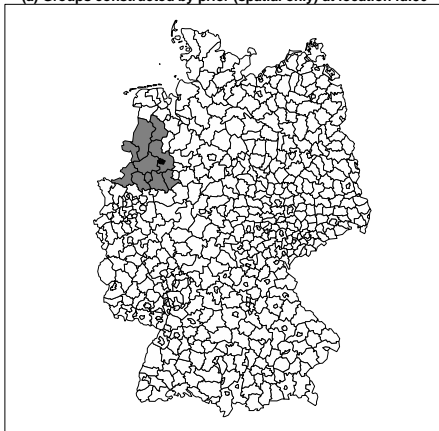
$$\begin{aligned} y_i | \eta_i &\sim \text{Poisson}(E_i \exp(\eta_i)) \\ \eta_i &= \mu + f_{rw}(x_i) + u_i + v_i, \end{aligned} \tag{5}$$

where μ is an intercept, \mathbf{u} is a spatially structured component, \mathbf{v} is an unstructured component and \mathbf{f}_{rw} is an intrinsic second-order random-walk model of the covariate x_i .



Groups

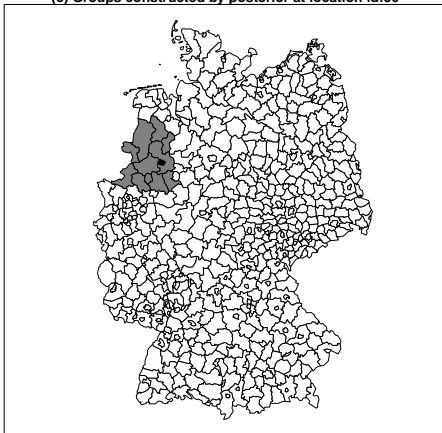
(a) Groups constructed by prior (spatial only) at location id:50





Groups

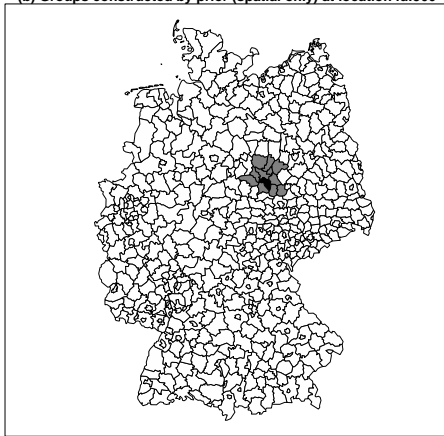
(c) Groups constructed by posterior at location id:50





Groups

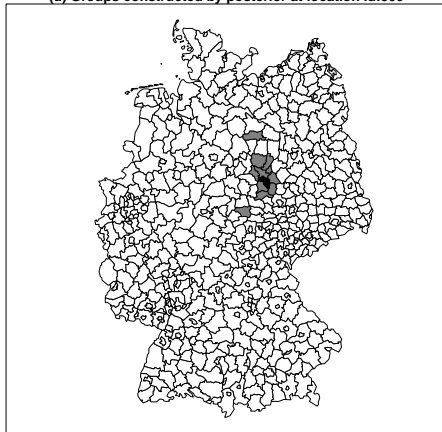
(b) Groups constructed by prior (spatial only) at location id:500





Groups

(d) Groups constructed by posterior at location id:500





Dengue risk in Brazil

The models study the influence of extreme hydrometeorological hazards on dengue risk, factoring in Brazil's urbanization levels. Our dataset, with 127, 224 samples representing 12, 895, 293 dengue cases, covers Brazil's 558 microregions from January 2001 to December 2019.

Data points include month, year, microregion, and state. The candidate covariates encompass the monthly average of daily minimum (T_{min}) and maximum temperatures (T_{max}), the palmer drought severity index (PDSI), the urbanization levels: overall (u), centered at high (u_1), intermediate (u_2), and more rural levels (u_3) and the access to water supply: overall (w) and centered at high-frequency shortages (w_1), intermediate (w_2), and low-frequency shortages (w_3).



Dengue risk in Brazil

The models study the influence of extreme hydrometeorological hazards on dengue risk, factoring in Brazil's urbanization levels. Our dataset, with 127, 224 samples representing 12, 895, 293 dengue cases, covers Brazil's 558 microregions from January 2001 to December 2019.

Data points include month, year, microregion, and state. The candidate covariates encompass the monthly average of daily minimum (T_{min}) and maximum temperatures (T_{max}), the palmer drought severity index (PDSI), the urbanization levels: overall (u), centered at high (u_1), intermediate (u_2), and more rural levels (u_3) and the access to water supply: overall (w) and centered at high-frequency shortages (w_1), intermediate (w_2), and low-frequency shortages (w_3).



Dengue risk in Brazil

The data generating model is chosen to be negative binomial, to account for overdispersion.

The latent field consists of a temporal component describing a state-specific seasonality using a cyclic first difference prior distribution and a spatial component describing year-specific spatially unstructured and structured random effects using a modified Besag-York-Mollie (BYM2) model with a scaled spatial component.

The temporal component has replications for each state, and the spatial component has replications for each year. We can express the base model using the INLA-style formula,

$$\begin{aligned}y(t, s) | \eta(t, s) &\sim \text{NB}(\mu(\eta(t, s)), \theta) \\ \eta(t, s) &= \beta_0 + \beta^\top \text{covariates} + f_t + f_s\end{aligned}$$

The number of parameters in this model is 21,567 with 127,224 observations for the full model.



Dengue risk in Brazil

The data generating model is chosen to be negative binomial, to account for overdispersion.

The latent field consists of a temporal component describing a state-specific seasonality using a cyclic first difference prior distribution

and a spatial component describing year-specific spatially unstructured and structured random effects using a modified Besag-York-Mollie (BYM2) model with a scaled spatial component.

The temporal component has replications for each state, and the spatial component has replications for each year. We can express the base model using the INLA-style formula,

$$\begin{aligned}y(t, s) | \eta(t, s) &\sim \text{NB}(\mu(\eta(t, s)), \theta) \\ \eta(t, s) &= \beta_0 + \beta^\top \text{covariates} + f_t + f_s\end{aligned}$$

The number of parameters in this model is 21,567 with 127,224 observations for the full model.



Dengue risk in Brazil

The data generating model is chosen to be negative binomial, to account for overdispersion.

The latent field consists of a temporal component describing a state-specific seasonality using a cyclic first difference prior distribution and a spatial component describing year-specific spatially unstructured and structured random effects using a modified Besag-York-Mollie (BYM2) model with a scaled spatial component.

The temporal component has replications for each state, and the spatial component has replications for each year. We can express the base model using the INLA-style formula,

$$\begin{aligned}y(t, s) | \eta(t, s) &\sim \text{NB}(\mu(\eta(t, s)), \theta) \\ \eta(t, s) &= \beta_0 + \beta^\top \text{covariates} + f_t + f_s\end{aligned}$$

The number of parameters in this model is 21,567 with 127,224 observations for the full model.



Dengue risk in Brazil

The data generating model is chosen to be negative binomial, to account for overdispersion.

The latent field consists of a temporal component describing a state-specific seasonality using a cyclic first difference prior distribution and a spatial component describing year-specific spatially unstructured and structured random effects using a modified Besag-York-Mollie (BYM2) model with a scaled spatial component.

The temporal component has replications for each state, and the spatial component has replications for each year. We can express the base model using the INLA-style formula,

$$\begin{aligned}y(t, s) | \eta(t, s) &\sim \text{NB}(\mu(\eta(t, s)), \theta) \\ \eta(t, s) &= \beta_0 + \beta^\top \text{covariates} + f_t + f_s\end{aligned}$$

The number of parameters in this model is 21,567 with 127,224 observations for the full model.



Dengue risk in Brazil

Index	Model	DIC	LOOCV	LGOCV		
				(m = 2)	(m = 3)	(m = 4)
1	$y \sim 1 + f_t + f_s$	3615.38	0.0151	0.0158	0.0206	0.0270
2	$y \sim 1 + T_{min} + f_t + f_s$	1562.96	0.0064	0.0067	0.0088	0.0098
3	$y \sim 1 + T_{max} + f_t + f_s$	2228.73	0.0091	0.0098	0.0133	0.0163
4	$y \sim 1 + PDSI + f_t + f_s$	2167.12	0.0092	0.0095	0.0126	0.0184
5	$y \sim 1 + PDSI + T_{min} + f_t + f_s$	160.43	0.0006	0.0006	0.0012	0.0023
6	$y \sim 1 + PDSI + T_{max} + f_t + f_s$	900.65	0.0038	0.0038	0.0057	0.0084
7	$y \sim 1 + PDSI + T_{min} + PDSI * u_1 + u + f_t + f_s$	38.21	0.0002	0*	0*	0*
8	$y \sim 1 + PDSI + T_{min} + PDSI * u_2 + u + f_t + f_s$	39.13	0.0002	0*	0*	0*
9	$y \sim 1 + PDSI + T_{min} + PDSI * u_3 + u + f_t + f_s$	28.64	0.0002	0*	0*	0*
10	$y \sim 1 + PDSI + T_{min} + PDSI * w_1 + w + f_t + f_s$	6.68	0*	0.0005	0*	0.0014
11	$y \sim 1 + PDSI + T_{min} + PDSI * w_2 + w + f_t + f_s$	0*	0*	0.0005	0*	0.0015
12	$y \sim 1 + PDSI + T_{min} + PDSI * w_3 + w + f_t + f_s$	4.55	0*	0.0006	0*	0.0014



Current work

- Point pattern model - log-Gaussian Cox Process model


$$Y_A | \lambda \sim \text{Poisson} \left(\int_A \lambda(s) ds \right), \quad \lambda(s) = \boldsymbol{\beta}^\top \mathbf{X} + u(s)$$

- Joint models - multiple endpoints for each person



Reference

- Liu, Z., van Niekerk, J. and Rue, H. “Leave-group-out cross-validation for latent Gaussian models”. SORT-Statistics and Operations Research Transactions, 2025; 49(1), pp. 121-146, doi:10.57645/20.8080.02.25.
- Van Niekerk, J., Krainski, E., Rustand, D. and Rue, H. A new avenue for Bayesian inference with INLA. Computational Statistics & Data Analysis, 2023; 181, pp. 107-135.



Thank you • شكرا



جامعة الملك عبد الله
للعلوم والتقنية
King Abdullah University of
Science and Technology