Fitting spatio-temporal geostatistical models in Stan using the bmstdr R package.

Sujit Sahu

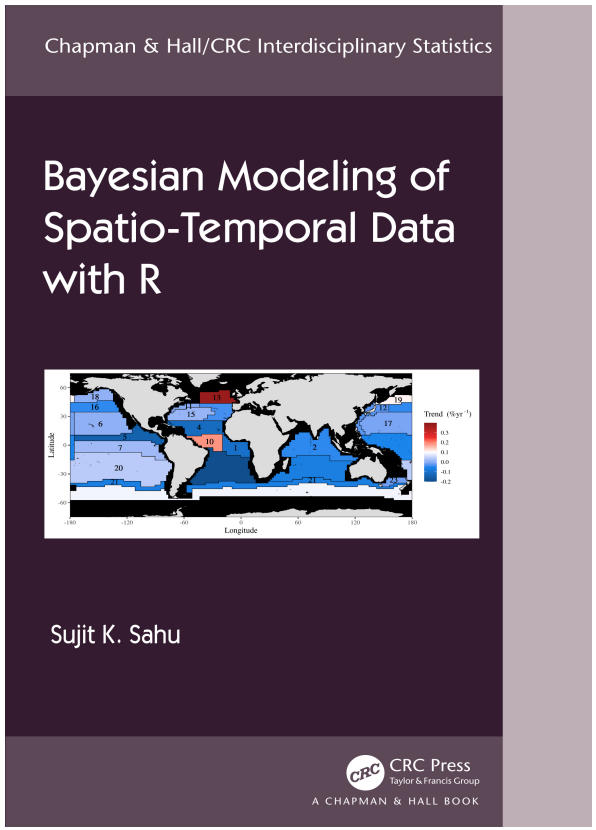
Email: S.K.Sahu@soton.ac.uk

Motivation

- There are many packages for performing spatio-temporal geostatistical modeling, e.g. INLA, spBayes, spTimer etc.
- Even the Stan user guide documents how to model using Gaussian process (GP) prior distributions. However, the illustrations are not general enough and only work for specific correlation functions, e.g. Gaussian.
- We are not aware of a Stan based package that can easily fit and validate GP based models along with other covariates.

Book BMSTDR and CRAN package bmstdr

- The bmstdr package accompanies the 2022 book "Bayesian Modeling of Spatio-Temporal Data with R."



- The full html version of the package vignette is on my site: https://www.sujitsahu.com

What's in this talk (outline)?

- It is only going to highlight what's in the package and the book.
- Mostly modeling point referenced, i.e. geostatistical data.
- I am not going to talk about how to program in stan!
- For that please see my book and also resources on the web.

What's on offer in the bmstdr CRAN package?

- For spatial and spatio-temporal modeling bmstdr offers three main modeling functions:
  1. bspatial: for spatial only data. Can fit using spBayes, INLA, rstan.
  2. bpstime: for spatio-temporal data. Can fit using spBayes, INLA, rstan, spTimer, sptDyn.
  3. Bmoving_sptime: for data from moving sensors.
  4. Bcartime: for areal data (both geospatial and temporal). Can fit using CARBayes, CARBayesST, INLA.
- All four functions work with S3 methods: summary, plot, residuals, fitted, coef and terms.
- Validations can be performed just by naming the rows of the model fitting data frame.

A general spatial model with nugget effect

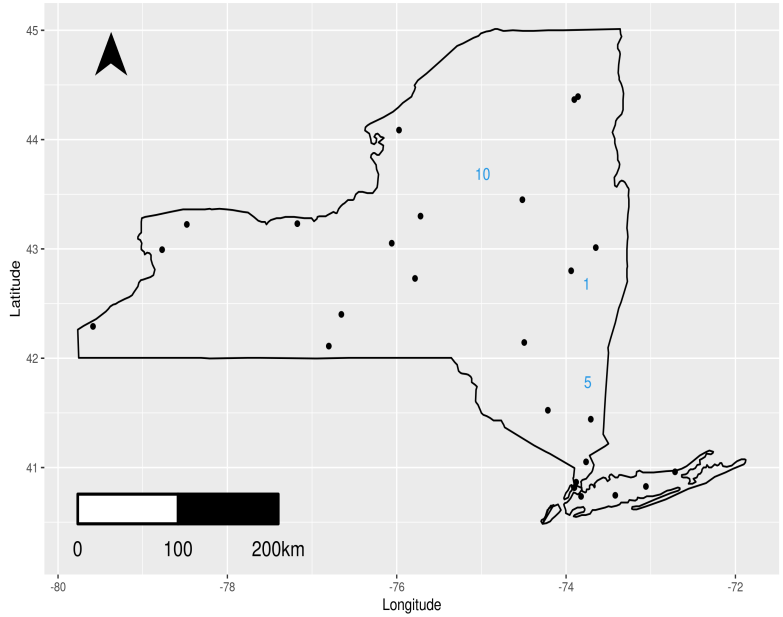


Figure 1: 28 air pollution monitoring sites in New York. This is a running example.

- Consider the general spatial model:

$$Y(\mathbf{s}_i) = \mathbf{x}'(\mathbf{s}_i)\beta + w(\mathbf{s}_i) + \epsilon(\mathbf{s}_i) \tag{1}$$

- The model has a linear regression part, a spatial component and an independent error.

bmstdr example code: Bspatial

```
f1 <- yo3~xmaxtemp+xwdsp+xrh
# Fit base (Bayesian) linear model
M1 <- Bspatial(formula=f1, data=nyspatial)
?Bspatial
# No nugget model
M2 <- Bspatial(model="spat", formula=f1, data=nyspatial,
        coordtype="utm", coords=4:5, phi=0.4)
# Marginal model using spBayes
M3 <- Bspatial(package="spBayes", formula=f1, data=nyspatial,
 coordtype="utm", coords=4:5, prior.phi=c(0.005, 2))
# Marginal model using spBayes
M4 <- Bspatial(package="stan", formula=f1, data=nyspatial,
coordtype="utm", coords=4:5,phi=0.4, mchoice=T)
print(M4); summary(M4); residuals(M4); plot(M4); coef(M4)
# Finally try INLA spde
M5  <- Bspatial(package="inla",formula=f1, data=nyspatial)
```

Model comparison results

Table 1: Model choice criteria for various models fitted to the nyspatial data set.

	M0	M1	M2	M3	M4	M5
pdic	2.07	4.99	4.98	5.17	5.31	4.17
pdicalt	13.58	5.17	5.16	7.83	6.46	NA
dic	169.20	158.36	158.06	158.68	158.75	157.23
dicalt	192.22	158.72	158.41	163.99	161.04	NA
pwaic1	1.82	5.20	4.93	4.88	4.93	4.73
pwaic2	2.52	6.32	5.91	6.77	5.96	NA
waic1	168.95	158.57	157.51	158.70	157.92	158.46
waic2	170.35	160.82	159.47	162.48	159.99	NA
gof	591.82	327.98	330.08	323.56	316.67	334.03
penalty	577.13	351.52	346.73	396.63	394.86	39.17
pmcc	1168.95	679.50	676.82	720.18	711.52	373.19

Model validation by passing the validrows argument.

```
s <- c(8,11,12,14,18,21,24,28)
M4.v  <- Bspatial(package="stan",formula=f1, data=nyspatial,
    coordtype="utm", coords=4:5,phi=0.4, validrows=s)
```

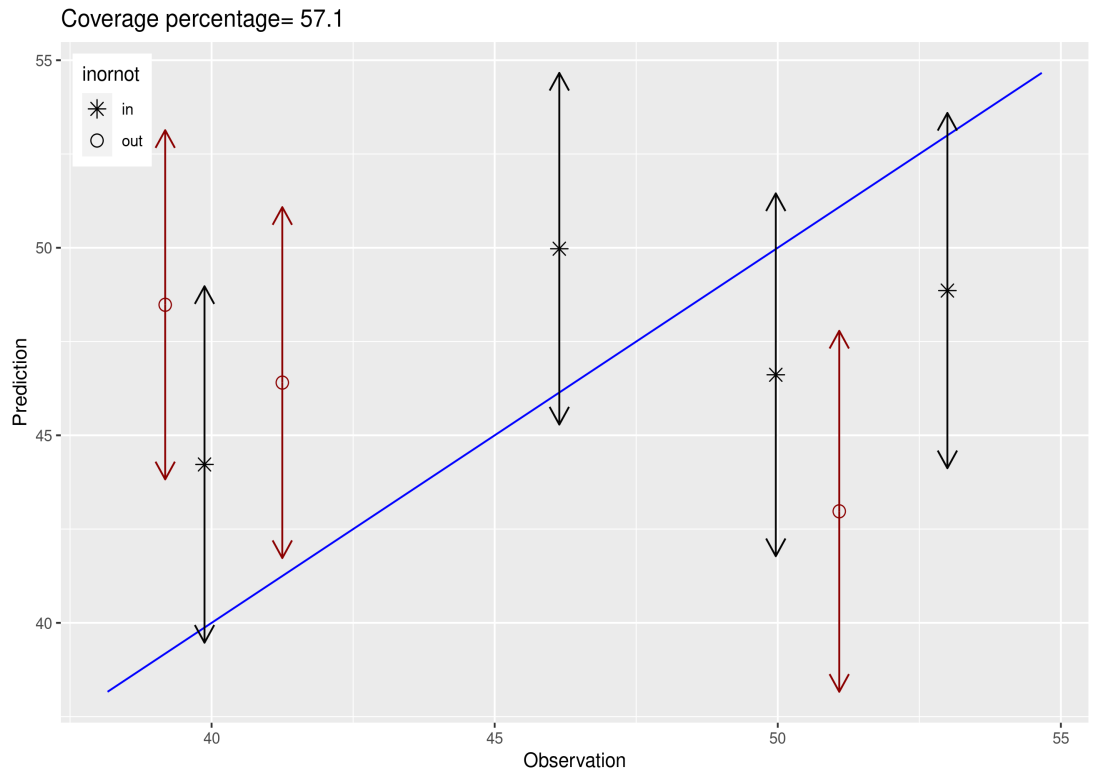


Figure 2: Prediction against observation plot with the prediction intervals included. The 'in/out' symbol in the plot indicates whether or not a prediction interval incudes the 45 degree line.

The Bsptime function for fitting spatio-temporal models

- The spatial model (1) is extended to the following spatio-temporal model.

$$Y(\mathbf{s}_i, t) = \mathbf{x}'(\mathbf{s}_i, t)\boldsymbol{\beta} + w(\mathbf{s}_i, t) + \epsilon(\mathbf{s}_i, t) \tag{2}$$

- Offers a variety of models: such as separable, auto-regressive and marginal modeling after integrating out the spatio-temporal component.
- Package options include stan, spTimer, spTDyn, spBayes, INLA.

Example code

```
f2 <- y8hrmax~xmaxtemp+xwdsp+xrh
M1 <- Bsptime(model="lm", formula=f2, data=nysptime,
              scale.transform = "SQRT")
M2 <- Bsptime(model="separable", formula=f2, data=nysptime,
    scale.transform = "SQRT", coordtype="utm", coords=4:5)
M3 <- Bsptime(package="spTimer", formula=f2, data=nysptime,
    model="GP", coordtype="utm", coords=4:5,
    scale.transform = "SQRT", N=N)
M4 <- Bsptime(package="stan",formula=f2, data=nysptime,
  coordtype="utm", coords=4:5, scale.transform = "SQRT",
  N=Nstan, burn.in=burn.in, mchoice=TRUE, verbose = F)
```

Model comparison results

Table 2: Model choice criteria for the four spatio-temporal models M1 to M4.

	M1	M2	M3	M4
pdic	4.98	4.98	78.65	30.36
pdicalt	4.94	4.95	841.96	31.22
dicorig	3912.07	3214.55	3132.10	2695.11
dicalt	3912.01	3214.50	4658.72	2696.83
pwaic1	4.85	14.39	48.53	9.05
pwaic2	4.87	14.58	132.90	10.04
waic1	3911.95	2448.00	2603.86	2088.15
waic2	3911.99	2448.38	2772.60	2090.12
gof	963.24	286.08	216.75	328.74
penalty	965.58	240.38	873.84	361.95
pmcc	1928.82	526.47	1090.59	690.69

Model validation results

- Set aside all the data from the eight sites as noted previously.
- This gives us 496 ($= 8 \times 62$) data points in the validation set and the model is fitted with remaining 1240 space-time observations.
- M1: Independent, M2: Separable, M3:spTimerGP, M4:stan, M5:INLA, M6:spTimerAR. M7:sptDyn, M9:spTimerGPP.

	M1	M2	M3	M4	M5	M6	M7	M9
rmse	9.35	6.49	6.40	6.42	9.73	6.46	6.59	6.36
mae	7.54	5.00	4.94	4.85	7.65	4.99	5.11	4.85
crps	5.67	10.56	6.79	3.23	2.64	5.97	5.12	7.47
cvg	98.36	99.59	99.59	92.62	65.16	99.39	99.39	99.39
gof	728.91	218.49	181.71	173.46	527.82	185.76	71.30	146.69
penalty	731.61	195.37	935.42	266.26	17.13	718.47	467.46	815.85
pmcc	1460.52	413.86	1117.13	439.72	544.95	904.23	538.76	962.54

- The M4:stan model is better!

Bmoving_sptime

Sahu and Challenor (2008) model ocean temperature data from the roaming Argo floats in the North Atlantic Ocean.

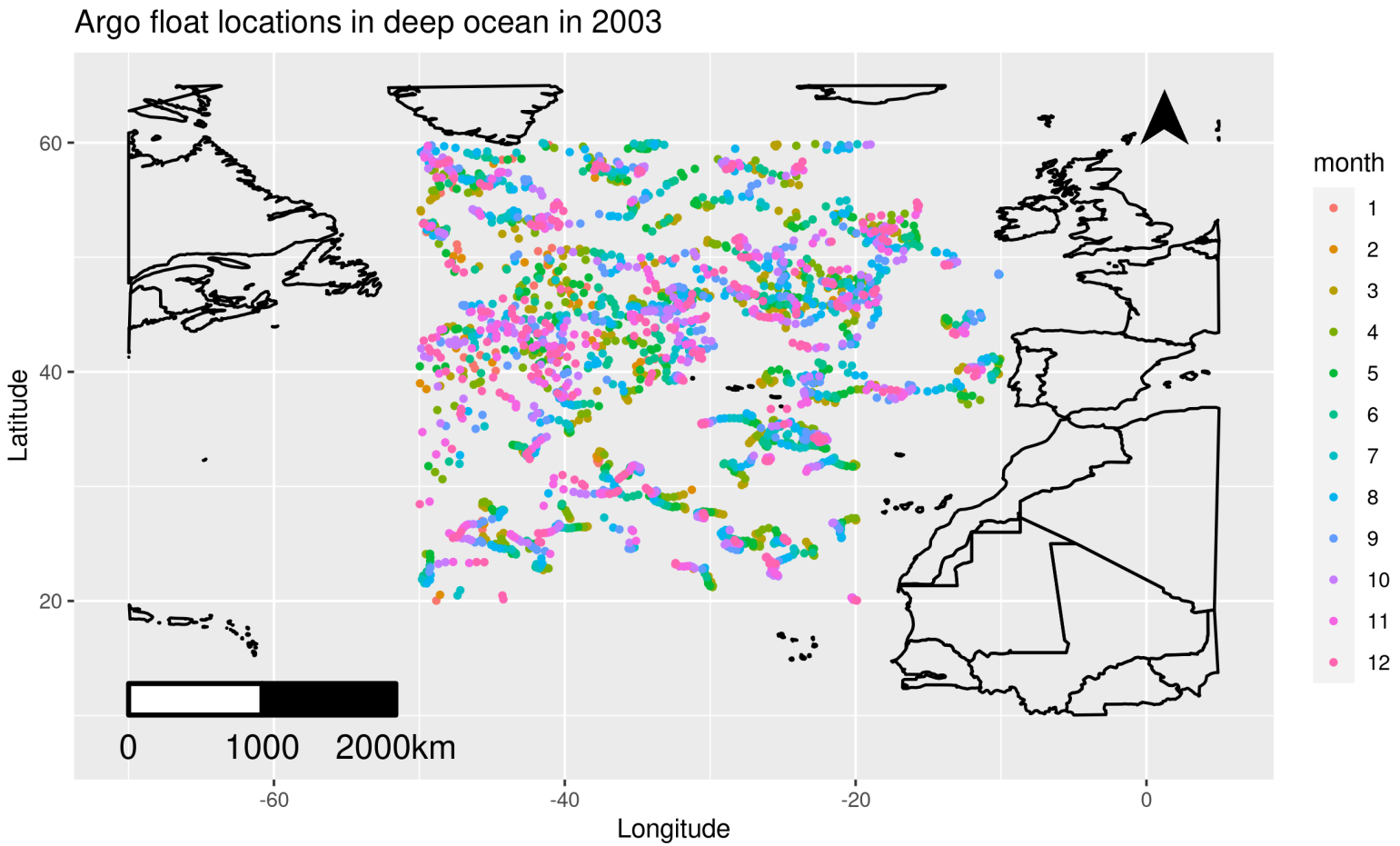


Figure 3: Locations of moving Argo floats in the deep ocean in 2003.

Marginal GP model for moving sensors

$$\mathbf{Y}_t \sim N\left(\mathbf{X}_t\beta, \quad \sigma_w^2 A_t S_w A_t' + \sigma_\epsilon^2 I\right), t = 1, \ldots, T \tag{3}$$

- $\mathbf{Y}_t$ and $\mathbf{X}_t$ are the vector of observations and covariate values at the n_t locations at time t.
- $A_t = C_t S_w^{-1}$, where S_w is $m \times m$ and has elements induced by the GP and C_t is $n_t \times m$ having the jth row and kth column entry.
- Further details for this model are provided in Section 8.6 in the book.
- The command below fits model (3) implemented by code written in Stan.

```
M2 <- Bmoving_sptime(formula=f2, data = deep,
  coordtype="lonlat", coords = 1:2, mchoice =TRUE)
```

Annual temperature map

- The model output can be processsed and predictions can be performed by writing additional code.
- See the full version of the vignette.
- Also the book website: https://www.sujitsahu.com/bookbmstdr/

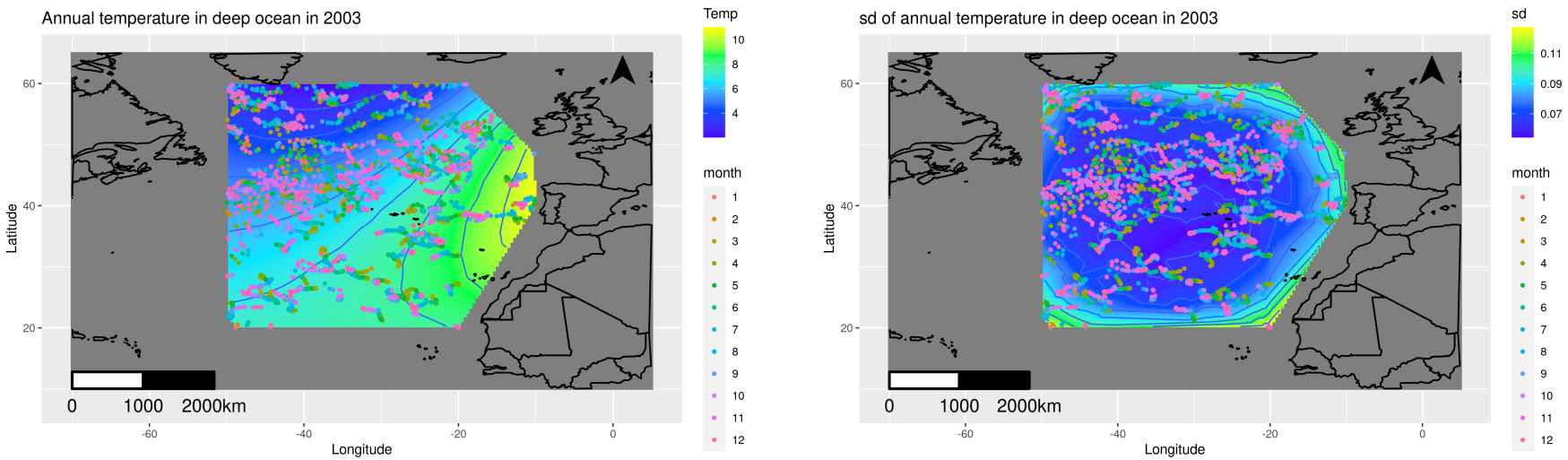


Figure 4: Annual prediction map (left panel) and sd of the predictions (right panel) of temperature at the deep ocean in 2003.

Bcartime is the bmstdr function for areal data modeling

- Fit using the CARBayes, CARBayesST, and INLA packages.
- Fit any Generalized Linear Model (GLM) with canonical link, e.g. logistic regression, Poisson log-linear and Gaussian Markov Random Fields.
- Including disease mapping and spatio-temporal disease mapping.
- Validate just by by naming the data rows
- Analyze using S3 methods such as summary and plot.

```
f1 <- highdeathsmr ~  jsa + log10(houseprice) + log(popdensity
M1st <- Bcartime(formula=f1, data=engdeaths, scol=scol,
tcol=tcol, trials=nweek, W=Weng, model="linear",
family="binomial", package="CARBayesST")
```

CAR modeling Illustration

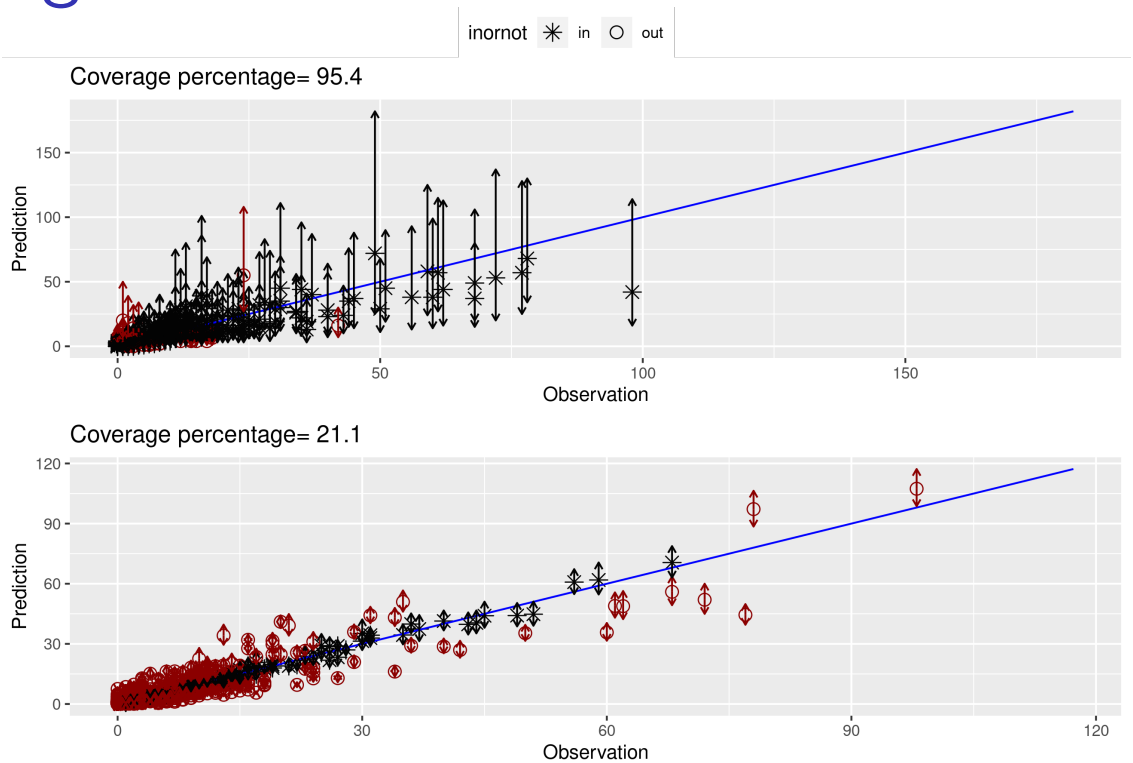


Figure 5: Predictions with 95% limits against observations for the AR (2) model fitted by CARBayesST package (left) and INLA (right).

- Coverage from the CARBayesST is about right.
- Experienced typically low coverage using INLA.
- Unfortunately no stan model for areal data.

Conclusion

- Stan provides the best model fits and out of sample validations.
- However, it may take longer to run and the user needs to write bespoke code.
- bmstdr provides four ready to use Stan implemented spatio-temporal geostatistical models.
- INLA users: Why not try Stan with the help of bmstdr? Stan users: Why not taste INLA just by calling bmstdr?
- For code and data go to my github repo: https://github.com/sujit-sahu
- My website https://www.sujitsahu.com/ provides a lot of resources.
- More needs to be done! I am looking for collaborators!!
- Please contact me if you have ideas to improve or if you have any feedback. S.K.Sahu@soton.ac.uk