Linking the SPDE method with Stan

Joaquin Cavieres

Universidad de Valparaíso, Chile
Ph.D (c) in Statistics

October 4, 2021

Contents

Motivation

Motivation

Sea urchin (*Loxechinus albus*) is one of the most important benthic resource in Chile ([6][15]). Due to their large-scale spatial metapopulation structure, sea urchin subpopulations are interconnected by larval dispersion, so the recovery of local abundance depends on the distance and hydrodynamic characteristics of their spatial domain. Besides, the population is structured as a metapopulation across a large spatial scale ([17] [18]).

Motivation

Figure: Common structure of the Sea urchin in a "patch". Reference: http://cocinafuturo.net/

....

Motivation

... and a set of "patches" form the sites of fishing.

Motivation

... and a set of "patches" form the sites of fishing.

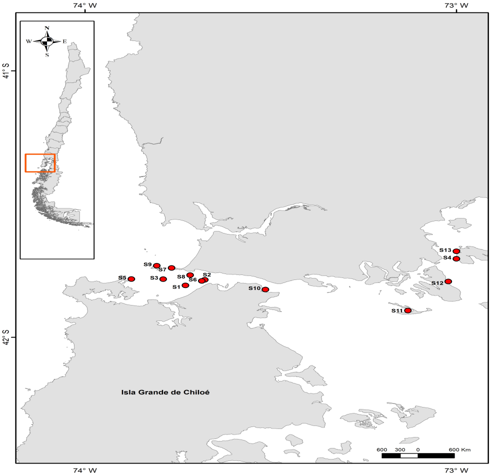


Figure: Sites of fishing of the Sea urchin (*Loxechinus albus*) in the south of Chile

So, what is the problem?

Motivation

So, what is the problem?

Currently, this resource is evaluated with classical stock assessment models, using standardized catch per unit effort (an index of relative abundance) as a key piece of information to determine catch quotas and achieve sustainability.

Motivation

So, what is the problem?

Currently, this resource is evaluated with classical stock assessment models, using standardized catch per unit effort (an index of relative abundance) as a key piece of information to determine catch quotas and achieve sustainability. However, these estimates assume hyperstability for the total population, ignoring spatial dependence among fishing sites, which is a fundamental concept for populations structured as metapopulation.

So, what is the problem?

Currently, this resource is evaluated with classical stock assessment models, using standardized catch per unit effort (an index of relative abundance) as a key piece of information to determine catch quotas and achieve sustainability. However, these estimates assume hyperstability for the total population, ignoring spatial dependence among fishing sites, which is a fundamental concept for populations structured as metapopulation.

.... and what is a standardized catch per unit effort?

Motivation

So, what is the problem?

Currently, this resource is evaluated with classical stock assessment models, using standardized catch per unit effort (an index of relative abundance) as a key piece of information to determine catch quotas and achieve sustainability. However, these estimates assume hyperstability for the total population, ignoring spatial dependence among fishing sites, which is a fundamental concept for populations structured as metapopulation.

.... and what is a standardized catch per unit effort?

Catch per unit effort (CPUE) is a crucial variable in fishery science and often assumed proportional to the abundance for a particular fishery resource over time

Motivation

Thus, we need to estimate the CPUE (index of relative abundance)!

Motivation

Thus, we need to estimate the CPUE (index of relative abundance)!

Methodologies

- Generalized Linear Models (GLM; [16][12])
- Generalized Additive Models (GAM's; [7][23][8])
- Generalized Linear Mixed Models (GLMM; [20][3])

Motivation

Recent methodologies applied

- Delta-GLMM ([22])
- TMB ([9])
- INLA ([4])

Hypothesis

Hypothesis

Hypothesis

- Incorporating the spatial dependence between sites of fishing (by larval dispersal) can improve the estimation of the CPUE ($\approx$ index of relative abundance)

Hypothesis

Hypothesis

- Incorporating the spatial dependence between sites of fishing (by larval dispersal) can improve the estimation of the CPUE ($\approx$ index of relative abundance)

Research objective

- Obtain the CPUE vector ($\approx$ index of relative abundance) incorporating the spatial and temporal dependence in the observations and compare it with the classical estimation.

Materials and Methods

Gaussian random field (GRF)

Definition

Let s a location in a particular area D and $\boldsymbol{u}(s)$ is a random effect (spatial) at that location. $\boldsymbol{u}(s) \in D$ is a stochastic processes and $\boldsymbol{D} \subset \mathbb{R}^d$ is the spatial domain where are measured the observations. $\boldsymbol{u}\,(s_i)$ is a realization of $\boldsymbol{u}(s)$ where $i = 1, \ldots, n$ locations. We asumme that $\boldsymbol{u}(s)$ has a multivariate Gaussian distribution (GRF), continuous over the space indexed by s and defined by it's mean and covariance ([5])

Note: We consider $s = (s_1, s_2)$ where s_1 and s_2 can be, for example, latitude and longitude respectively.

Materials and Methods

We can assume a GRF for the spatial underlying process to model geostatistical data but it's very expensive in computational terms

Materials and Methods

We can assume a GRF for the spatial underlying process to model geostatistical data but it's very expensive in computational terms

Solution?

We can assume a GRF for the spatial underlying process to model geostatistical data but it's very expensive in computational terms

Solution?

An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach

Finn Lindgren and Håvard Rue

Norwegian University of Science and Technology, Trondheim, Norway

and Johan Lindström

Lund University, Sweden

[11] propose an approximation of the GRF by a GMRF using Stochastic Partial Differential Equations (SPDE)

Materials and Methods

More references about the SPDE method can be found in:

Bayesian Spatial Modelling with R-INLA

Finn Lindgren
University of Bath, United Kingdom

Håvard Rue
Norwegian University of Science and Technology, Norway

Understanding the stochastic partial differential equation approach to smoothing

David L Miller * Richard Glennie* Andrew E Seaton*

Figure: Literature to read about the SPDE method

[10] , [13]

So, how can we express that in a statistical model?

So, how can we express that in a statistical model? $\rightarrow$ Hierarchical model!

So, how can we express that in a statistical model? $\rightarrow$ Hierarchical model!

$$
\begin{aligned}
\boldsymbol{\theta} &\sim \boldsymbol{\theta} && \text{Hyperparameters} && (1)\\
\boldsymbol{u} \mid \boldsymbol{\theta} &\sim \mathscr{N}(0, \boldsymbol{Q}(\boldsymbol{\theta})^{-1}) && \text{Latent gaussian field} && (2)\\
y_i \mid \boldsymbol{u}, \boldsymbol{\theta} &\sim \prod_i \pi(y_i \mid \eta_i, \boldsymbol{\theta}) && \text{Observations} && (3)
\end{aligned}
$$

where $\boldsymbol{Q}(\boldsymbol{\theta})$ is the presicion matrix, $\boldsymbol{u}$ is the latent gaussian field and $\eta_i = log(\mu_i) = \text{intercept} + f(\boldsymbol{X}_i) + \boldsymbol{u}_i$, where the matrix $\boldsymbol{X}$ is a set of covariates and $\boldsymbol{u} \sim GMRF(0, \boldsymbol{Q}^{-1})$

How to solve in an efficient way a (spatial) Hierarchical Bayesian model?

Materials and Methods

How to solve in an efficient way a (spatial) Hierarchical Bayesian model?

- `R-INLA` $\Longrightarrow$ ([21], [2], [10], [1])

Materials and Methods

How to solve in an efficient way a (spatial) Hierarchical Bayesian model?

- `R-INLA` $\Longrightarrow$ ([21], [2], [10], [1])
- MCMC ... `Stan`? `How`?

Materials and Methods

Template Model Builder (TMB)

TMB is a frequentist software/package for fitting statistical latent variable models to the data.

The TMB code is essentially written in C++ and there is no need to supply the derivatives of the function to be minimized with respect to the parameters; these are computed automatically using Automatic differentiation.

Materials and Methods

Also,

- The sparsity of the Hessian is detected automatically.
- Automatic bias-corrections are applied when reporting a non-linear function of the random effects.
- The model can be parallelized.

Materials and Methods

How works `TMB`?

- At first intance we need to define the log-likelihood:

$$f(\theta, u) = log(Pr(y \mid \theta_1, u)Pr(u \mid \theta_2)))$$

- Laplace approximation (Taylor series expansion of second order) for the joint log-likelihood:

$$f(u, \theta) \approx f(\hat{u} \mid \theta) + f^{'}(\hat{u} \mid \theta)(\hat{u} - u) + 1/2 * f^{''}(\hat{u} \mid \theta)(\hat{u} - u)^2$$

- Evaluate the Laplace approximation around "inner maximum"

$$\hat{u} = \underset{u}{\operatorname{argmax}}(f(\theta, u))$$

- Aproximate the joint likelihood

$$Pr(y \mid \theta_1, u)Pr(u \mid \theta_2) = e^{f(u|\theta)} \approx e^{f(\hat{u}|\theta) - 1/2 * |f^{''}(\hat{u})|(\hat{u} - u)^2}$$

Materials and Methods

If you want to know more about spatial modeling in `TMB` you can read this very interesting and applied article (it has an example code written in `TMB`!):

A Statistical Introduction to Template Model Builder:
A Flexible Tool for Spatial Modeling

Aaron Osgood-Zimmerman[a,*], Jon Wakefield[a,b]

University of Washington, Seattle

[a] *Department of Statistics*
[b] *Department of Biostatistics*

[19]

Materials and Methods

Ok, but where `Stan` appear?

Ok, but where `Stan` appear? $\Longrightarrow$ `tmbstan` [14]

Ok, but where Stan appear? $\Longrightarrow$ tmbstan [14]

tmbstan

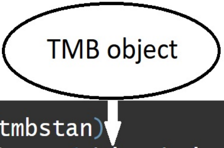
The models built in TMB pass their log-density and gradients calculations to the Bayesian samplers in Stan through a R interface. You only need to add the prior distribution on the parameters in your TMB code!

Materials and Methods

Ok, but where `Stan` appear? $\Longrightarrow$ `tmbstan` [14]

`tmbstan`

The models built in `TMB` pass their log-density and gradients calculations to the Bayesian samplers in `Stan` through a R interface. You only need to add the prior distribution on the parameters in your `TMB` code!




```
library(tmbstan)
fit = tmbstan(Obj, chains = 2, open_progress = FALSE,
              init ='last.par.best',
              control = list(max_treedepth = 10, adapt_delta = 0.8)
              iter = 2000, warmup=1000, seed=483892929)
```

CPUE estimation

Data

- Temporal observations: from 1996 to 2016 ("Year" variable as a "factor").
- Spatial observations: 13 sites of fishing ("sites" as a spatial random effect)
- Covariates: "Depth" (average depth of catches), "Quarter" (season of the year), and the variable "Market" (1 or 2).

CPUE estimation

Table: Proposed models to obtain a relative abundance index

Models	Structure
Lognormal	$(y_i \mid \boldsymbol{\theta}) \sim p(y_i \mid \eta_i, \boldsymbol{\theta})$
Spatial Lognormal	$(y_i \mid \boldsymbol{u}, \boldsymbol{\theta}) \sim p(y_i \mid \eta_i, \boldsymbol{\theta})$
Gamma	$(y_i \mid \boldsymbol{\theta}) \sim p(y_i \mid \eta_i, \boldsymbol{\theta})$
Spatial Gamma	$(y_i \mid \boldsymbol{u}, \boldsymbol{\theta}) \sim p(y_i \mid \eta_i, \boldsymbol{\theta})$

Results

Results

Table: Comparison with `LOO` criterion for each proposed model. The `elpd_diff` measures the difference between each model relative to the best $\widehat{elpd}_n$ (the model in the first row) and `se_diff` is the standard error of the difference in `elpd_diff`.

Models	elpd_diff	se_diff
Spatial Gamma	0	0
Spatial Lognormal	-42	23
Lognormal	-88	18
Gamma	-125	28

Results

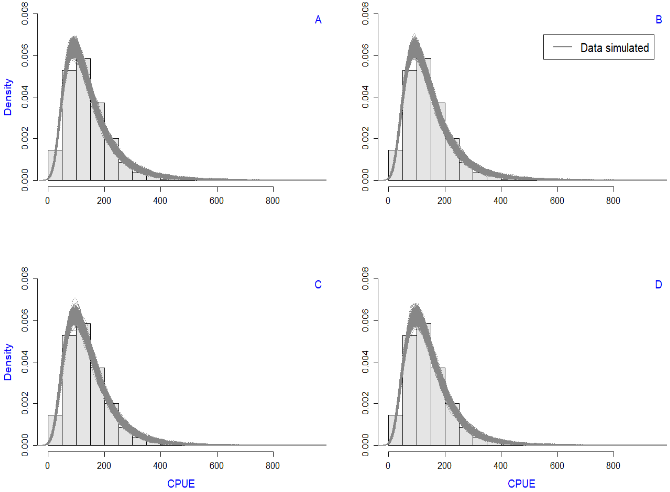


Figure: Observed data (histogram) and density estimates (lines) of 1,000 posterior predictive data simulated from Lognormal model, spatial Lognormal model, Gamma model and spatial Gamma model. See Table 1 for model descriptions.

Results

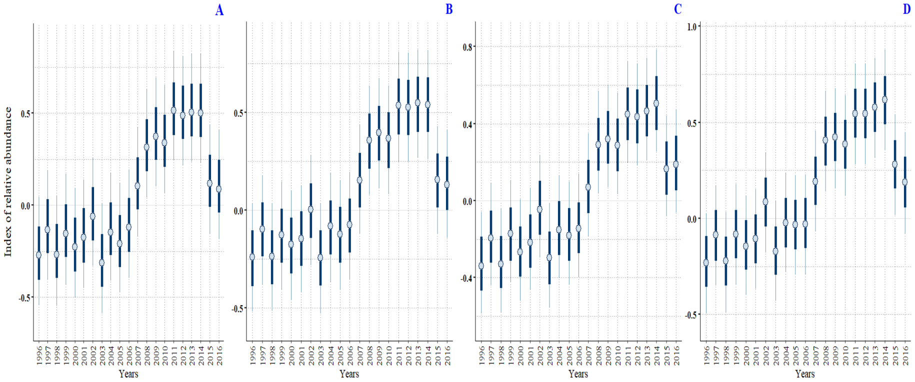


Figure: Comparison of the coefficients (relative abundance indices) estimated with Lognormal model, spatial Lognormal model, Gamma model and spatial Gamma model. The points are the values for the coefficients and the thick bars are uncertainty intervals computed from posterior draws with all chains merged in the MCMC method (90% credible interval).

Results

To assess the potential effects of including sites with only one year of observations we made two additional model comparisons.

Table: Additional comparisons with LOO

Comparison excluding site 1		
Models	elpd_diff	se_diff
Spatial Gamma	0	0
Spatial Lognormal	-15	23
Gamma	-89	18
Lognormal	-95	28

Comparison excluding sites 1, 2, 3 and 8		
Models	elpd_diff	se_diff
Spatial Lognormal	0	0
Spatial Gamma	-5	22
Lognormal	-70	16
Gamma	-81	28

Results

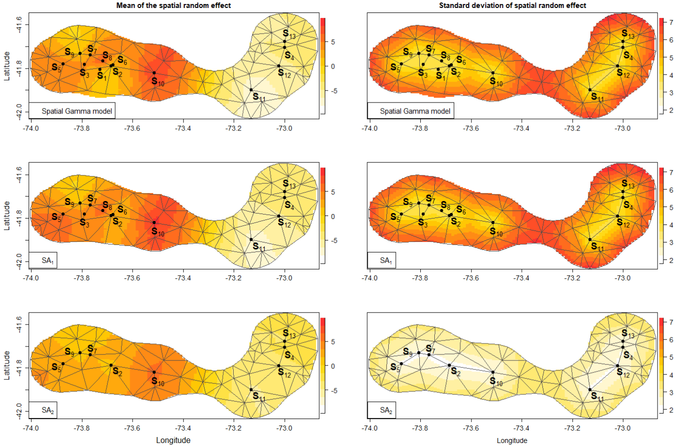


Figure: Mean (left) and standard deviation (right) of the spatial field estimated by the spatial Gamma model, and for the first (SA_1) and second (SA_2) sensitivity analysis respectively.

Conclusions

Conclusions

- Incorporating a spatial random effect to obtain a relative abundance index for sea urchin (L. albus) enabled better statistical performance than models without spatial dependence.
- Although the trends of the estimated indices for our case study with and without spatial effects were similar, statistical diagnostics clearly indicated that the spatial model outperformed the non-spatial version and fit the data better.
- This difference could have important impacts on the estimated status and trend of the stock, and ultimately the catch quota, so assessing the stock with both indices would be valuable.

Acknowledgments

- This research was done at Aalto University (Internship) supervised by Aki Vehtari
- We had a valuable help of Cole C Monnahan (NOAA)

ELSEVIER

Contents lists available at ScienceDirect

Fisheries Research

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

Check for updates

Accounting for spatial dependence improves relative abundance estimates in a benthic marine species structured as a metapopulation

Joaquin Cavieres[a,*], Cole C. Monnahan[b], Aki Vehtari[c]

[a] Instituto de Estadística, Facultad de Ciencias, Universidad de Valparaíso, Valparaíso, Chile
[b] Resource Ecology and Fisheries Management, National Marine Fisheries Service (NOAA), Seattle, WA, United States
[c] Department of Computer Science, Aalto University, Finland

References I

H. Bakka, H. Rue, G.-A. Fuglstad, A. Riebler, D. Bolin, J. Illian, E. Krainski, D. Simpson, and F. Lindgren.
Spatial modeling with r-inla: A review.
Wiley Interdisciplinary Reviews: Computational Statistics, 10(6):e1443, 2018.

M. Blangiardo, M. Cameletti, G. Baio, and H. Rue.
Spatial and spatio-temporal models with r-inla.
Spatial and spatio-temporal epidemiology, 4:33–49, 2013.

B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White.
Generalized linear mixed models: a practical guide for ecology and evolution.
Trends in ecology & evolution, 24(3):127–135, 2009.

References II

J. Cavieres and O. Nicolis.
Using a spatio-temporal bayesian approach to estimate the relative abundance index of yellow squat lobster (cervimunida johni) off chile.
Fisheries research, 208:97–104, 2018.

N. A. Cressie.
Spatial prediction and kriging.
Statistics for Spatial Data (Cressie NAC, ed). New York: John Wiley & Sons, pages 105–209, 1993.

C. Guisado.
Historia de vida, reproducción y avances en el cultivo del erizo comestible chileno l. albus (molina, 1782)(echinoidea; echinidae).
Manejo y desarrollo pesquero, pages 59–68, 1987.

References III

T. Hastie and R. Tibshirani.
Generalized additive models: some applications.
Journal of the American Statistical Association, 82(398):371–386, 1987.

T. J. Hastie and R. J. Tibshirani.
Generalized additive models.
Routledge, 2017.

K. Kristensen, A. Nielsen, C. W. Berg, H. Skaug, and B. Bell.
Tmb: automatic differentiation and laplace approximation.
arXiv preprint arXiv:1509.00660, 2015.

F. Lindgren, H. Rue, et al.
Bayesian spatial modelling with r-inla.
Journal of Statistical Software, 63(19):1–25, 2015.

References IV

F. Lindgren, H. Rue, and J. Lindström.
An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(4):423–498, 2011.

P. McCullagh and J. A. Nelder.
Generalized linear models.
Routledge, 2019.

D. L. Miller, R. Glennie, and A. E. Seaton.
Understanding the stochastic partial differential equation approach to smoothing.
Journal of Agricultural, Biological and Environmental Statistics, 25(1):1–16, 2020.

References V

C. C. Monnahan and K. Kristensen.
No-u-turn sampling for fast bayesian inference in admb and tmb: Introducing the adnuts and tmbstan r packages.
PloS one, 13(5):e0197954, 2018.

C. A. Moreno, C. Molinet, P. Díaz, M. Díaz, J. Codjambassis, and A. Arévalo.
Bathymetric distribution of the chilean red sea urchin (loxechinus albus, molina) in the inner seas of northwest patagonia: Implications for management.
Fisheries Research, 110(2):305–311, 2011.

J. A. Nelder and R. W. Wedderburn.
Generalized linear models.
Journal of the Royal Statistical Society: Series A (General), 135(3):370–384, 1972.

References VI

J. Orensanz and G. S. Jamieson.
The assessment and management of spatially structured stocks: an overview of the north pacific symposium on invertebrate stock assessment and management.
Canadian Special Publication of Fisheries and Aquatic Sciences, pages 441–459, 1998.

J. Orensanz, M. Pascual, and M. Fernández.
Biology and fisheries of the scallops from the southwest atlantic ocean.
Scallops: biology, ecology and aquaculture, pages 981–999, 1991.

A. Osgood-Zimmerman and J. Wakefield.
A statistical introduction to template model builder: A flexible tool for spatial modeling.
arXiv preprint arXiv:2103.09929, 2021.

References VII

J. C. Pinheiro and D. M. Bates.
Linear mixed-effects models: basic concepts and examples.
Mixed-effects models in S and S-Plus, pages 3–56, 2000.

H. Rue, S. Martino, and N. Chopin.
Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations.
Journal of the royal statistical society: Series b (statistical methodology), 71(2):319–392, 2009.

J. T. Thorson and E. J. Ward.
Accounting for vessel effects when standardizing catch rates from cooperative surveys.
Fisheries Research, 155:168–176, 2014.

References VIII

S. Wood and M. S. Wood.
Package ‘mgcv’.
R package version, 1:29, 2015.

Thank You