A poorman's spatio-temporal gridded HBM to handle 0-inflated biomass data

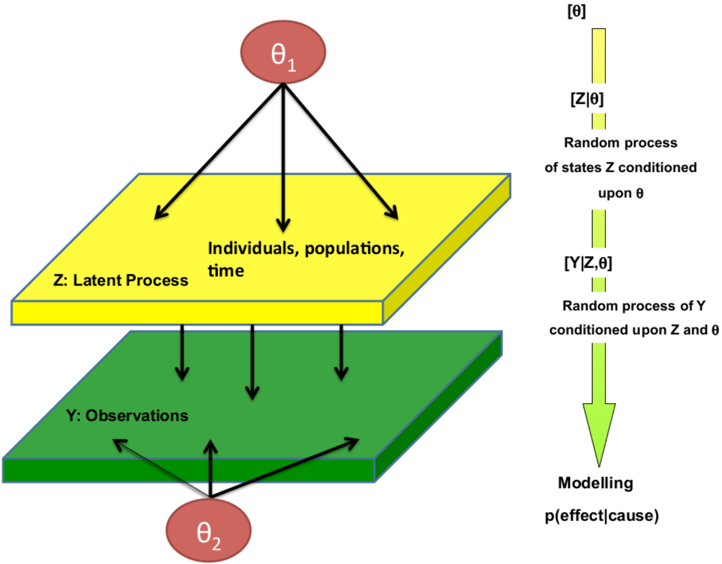
Jean-Baptiste Lecomte, Éric Parent, Liliane Bel, Marie P. Etienne

AgroParisTech

Contributor: Hugues Benoit(Environnement Canada, Moncton)

Mai 16, 2017

Observation/Latent/Parameters



An up-to-date methodological impulse towards highly sophisticated spatio-temporal frameworks

The sandwich layer has grown uneatable

Are things getting nowadays overbig?

- sample size grows: Records of environmental data to be indexed with s, t
- model sophistication increases : Environmental data more alike when closer in time and space.

HBM as an attempt to distangle complexity:

$$\theta \sim \pi(\theta) \text{ parameters}$$

$$Z|\theta \sim g(z|\theta) \quad \text{latent var.}$$

$$Y|z,\theta \sim \prod_i^n f(y_i|z_i,\theta) \text{ data}$$

Model and Inference challenges although workable multivariate process model: $g(.)$ often normal with $\mathbb{E}(Z|\theta) = \mu_\theta(.), \mathbb{V}(Z|\theta) = \Sigma_\theta(.,.)$

Some good and bad reasons to keep it simple

Making compromises is unavoidable

Two possible strategies:

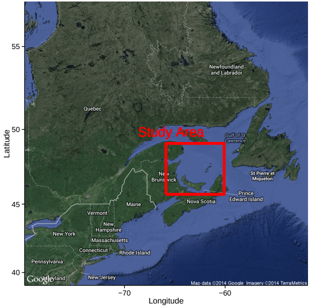
1. Simplify the model:
   - approximate with a Gaussian Markov Random Field (Lindgren et al., 2011) requiring $O(nlogn)$ operations
   - work out low rank methods (Cressie and Johannesson, 2008) $O(n)$
2. Keep the model but forget about the "golden standards" for inference:
   - Frequentists: pseudo, partial, composite, (...)-likelihoods e.g. Curriero and Lele (1999)
   - Bayesians: INLA, Variational Bayes, etc.

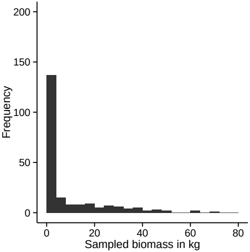
Basic requirements

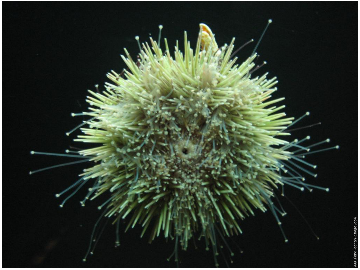
A minima

- Favoring a straight interface with others, possible non aware of statistical modeling
- Allowing for some increase of the database,
- Observation submodel may be specific,
- Trade-off between biologic realism and model complexity, isn't the best inference procedure the one you know?

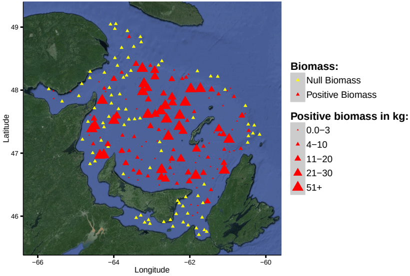
Urchins in the south Saint-Lawrence Gulf

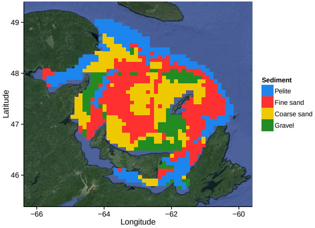


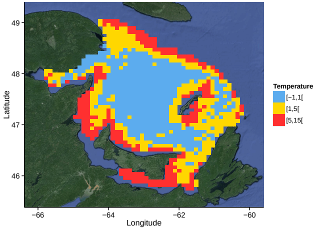


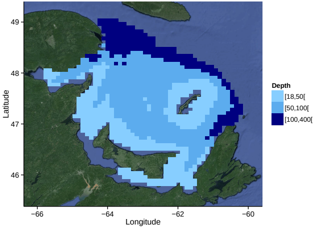


Urchins in the south Saint-Lawrence Gulf (1997)

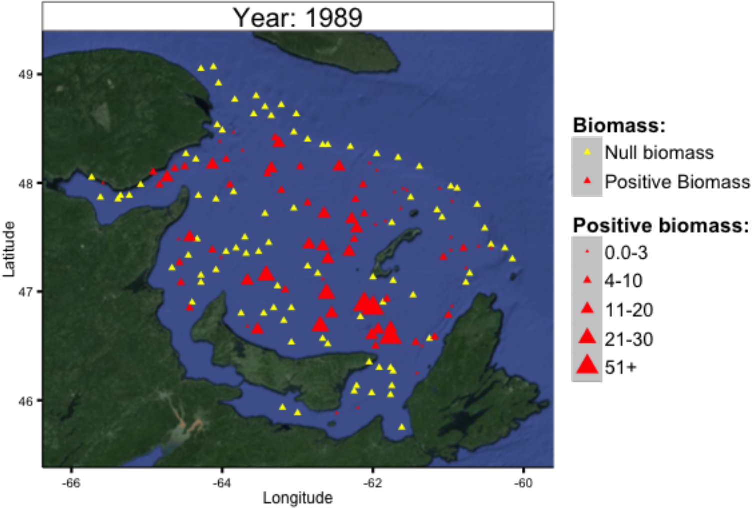




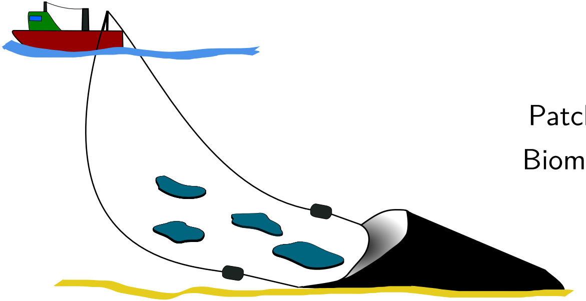




Spatio-temporal data



Observation process: A compound Poisson Gamma model



s site, $i \in s$

$$\begin{aligned} \text{Patches:} \quad & N_s \sim \text{Poisson}(\lambda_s) \\ \text{Biomass:} \quad & M_{s,i} \sim \text{Gamma}(a_s, b_s) \\ Y_s \quad & = \sum_{i=1}^{N_s} M_{s,i} \end{aligned}$$

(λ_s, a_s, b_s) are latent variables stemming from the (spatio-temporal) process layer.

Y_s takes 0 when $N_s = 0$

Tweedie model not in INLA

Process Convolution: from Random fields to Mixed effects

If $X(s)$ continuous white noise over domain S, $k(.)$ smoothing kernel:

$$Z(s) = \int_{u \in S} X(u)k(u-s)du$$

$\Delta = s - s'$

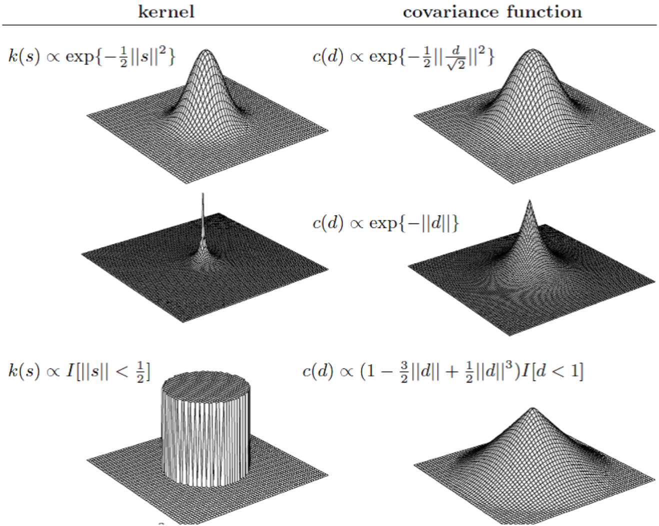
$$c(d) = cov(Z(s), Z(s')) = \int_{u \in S} k(u)k(u-d)du$$

Z isotropic : The spectrum of $c(.)$ is the square of the Fourier transform of $k(.)$

Space-time extension

$$Z(s,t) = \int_{u \in S} X(u,t)k(u-s)du$$

Process Convolution: from Random fields to Mixed effects



Process Convolution: from Random fields to Mixed effects

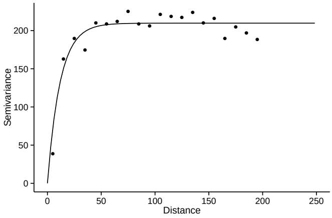
Process convolution approach for space+AR for time
(patch nb $\lambda(s,t) \leftarrow Z(s,t)$+Covariates)

$$Z(s,t) \approx \sum_{g=1}^{G} X_g(t) K(g,s)$$

$$X(t+1) = \rho X(t) + \varepsilon(t)$$

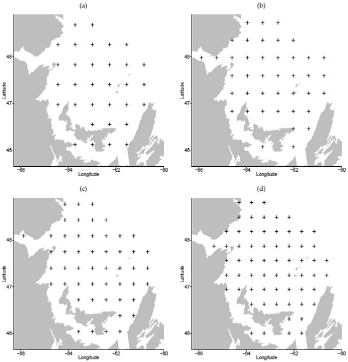
$$\varepsilon \sim N(0, \sigma^2 I)$$

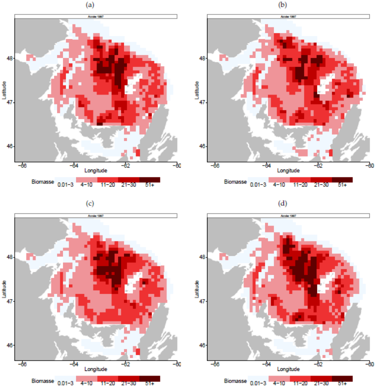
$$K(g,s) = \exp - \left(\frac{d(s,g)}{\phi} \right)$$



(G grid points)

Does the grid size matter?





Nb points	32	41	50	63
DIC	1155.7	1146.2	1137.8	1175.7

Inference

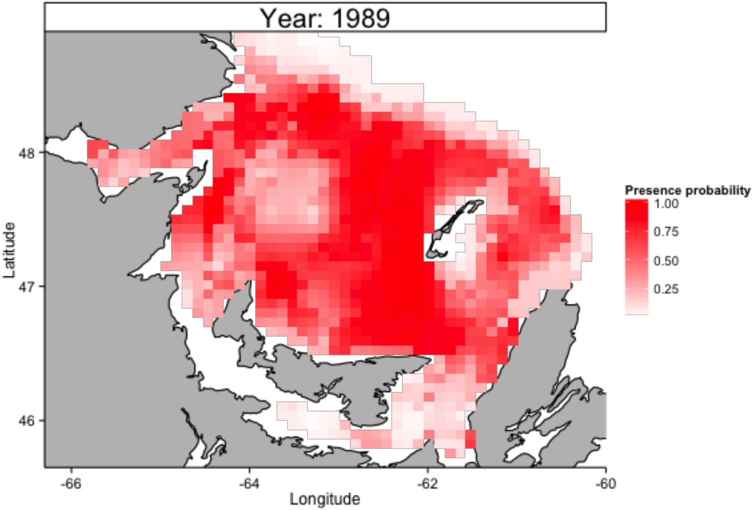
- Mixed-effect hierarchical model,
- Ready-made framework,
- Launch compiled BUGS code from NIMBLE.

Friendly softwares, e.g. WinBUGS®

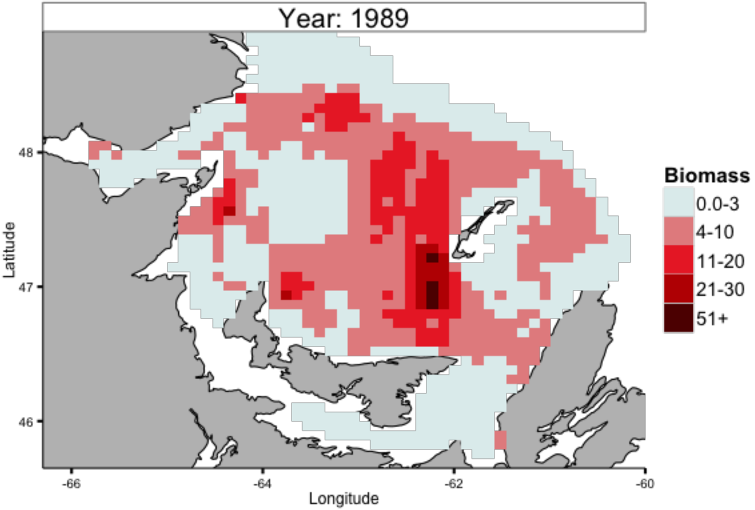
"Bayesian inference Using Gibbs Sampler"
Medical Research Council, Biostatistics Unit, Cambridge, UK
Imperial College, London, UK

- **Bayesian statistical modeling**
- **Very popular**
- **Sampling based methods using Gibbs sampling (eventually hybrid)**
- **Models can be described graphically**
- **Freeware, available at : www.mrc-bsu.cam.ac.uk/bugs/**
- **Extensions :** { **Spatial models : GeoBUGS®** / **Convergence : Coda®, Boa® (R packages)** }
- **Coupling with R**

Mean posterior presence probability



Mean posterior expected biomass



Conclusion

- Low cost spatio-temporal approach,
- Misalignment is no longer an issue,
- Model is not anymore curved by the sample locations,
- Time and space are modeled one after another.

Perspectives

- Computational burden : 3 hours!
- Model with more realistic ecological transition,
- Selection of explanatory variables,
- Irregular grids, special grid location for zeros,
- Consider the grid center locations as an unknown of the model,
- Pushing to higher dimensions (Multivariate extensions for Spatio-temporal Predator/prey models).

Thanks for your attention !

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 70:825–848.

Calder, C. A. (2003). *Exploring Latent Structure in Spatial Temporal Processes Using Process Convolutions*. PhD thesis, Duke University.

Calder, C. A. (2007). Dynamic factor process convolution models for multivariate space-time data with application to air quality assessment. *Environmental and Ecological Statistics*, 14(3):229–247.

Cressie, N. (1993). *Statistics for Spatial Data, revised edition*. Wiley, New York.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society, Series B*, 70:209–226.

Curriero, F. and Lele, S. (1999). A composite likelihood approach to semivariogram estimation. *Journal of Agricultural, Biological and Environmental Statistics*, 4:9–28.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics*, 5(2):173–190.

Higdon, D. (2002). Space and space-time modeling using process convolutions. In *Quantitative methods for current environmental issues*, pages 37–56. Springer Verlag, c. anderso edition.

Hooten, M. B. and Wikle, C. K. (2007). A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove. *Environmental and Ecological Statistics*, 15(1):59–70.

Kern, J. (2000). *Bayesian process-convolution approaches to specifying spatial dependence structure*. PhD thesis, Duke University.

Lecomte, J.-B., Benoît, H. P., Ancelet, S., Etienne, M.-P., Bel, L., and Parent, E. (2013). Compound Poisson-gamma vs. delta-gamma to handle zero-inflated continuous data under a variable sampling volume. *Methods in Ecology and Evolution*, 4(12):1159–1166.

Lindgren, F., Rue, H., and Lindstrom, J. (2011). 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*, 73:423–498.

Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*, Wiley, Hoboken, NJ.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B*, 79:319–392.