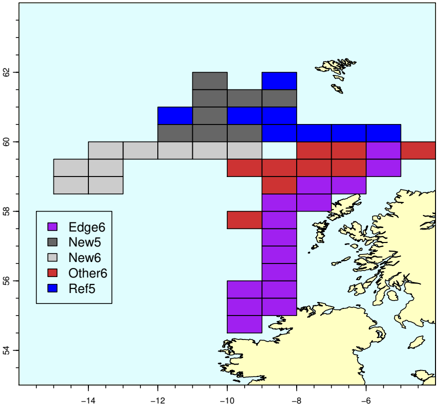
Using generalised additive mixed models for estimating relative abundance of blue ling for fisheries stock management

Nicole Augustin,[1] **Verena Trenkel**[2]**, Pascal Lorance**[2] **& Simon Wood**[3]
[1]august consulting, Bath, UK; [2]IFREMER, Nantes, France; [3]University of Bristol, UK

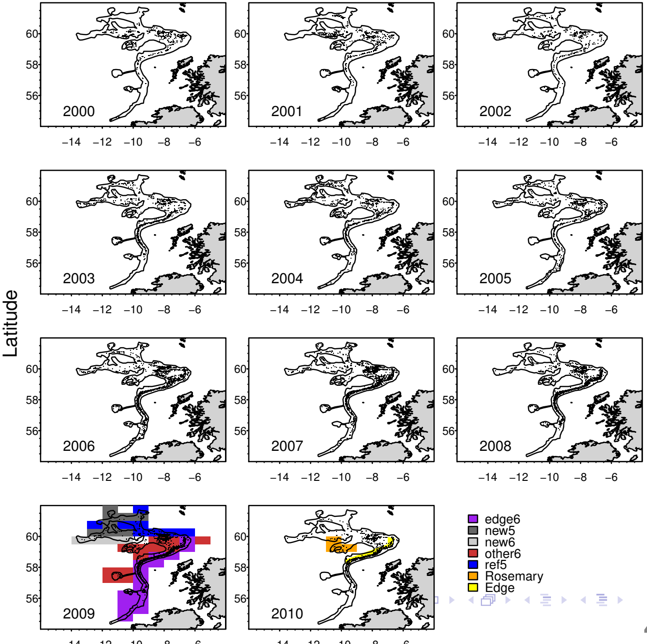
Blue Ling Area of Interest



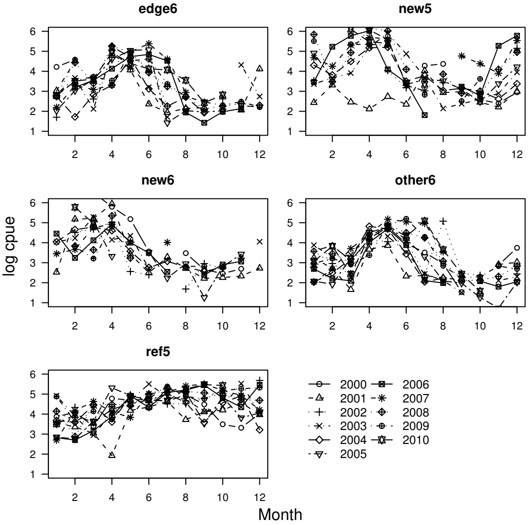
The Data

- 19 French deep-sea trawlers operating in the Northeast Atlantic during the period 2000-2010.
- Variables: landings (biomass in kg) by species, latitude and longitude, mean fishing depth, haul duration a proxy for effort .
- Use subset: haul duration between 30 and 600 mins, haul depth between 200 and 1100m
- Zero landings indicate no abundance or very low abundance of blue ling in the specific area and time.
- By-catch of blue ling is always possible (not affected by differences in fishing techniques due to targetting).

Positions of hauls per year



Observed median monthly catch (log kg) per hour by year and fishing areas.



The questions

- What is the relative overall trend of blue ling abundance?
- Is there any evidence for a space-time interaction, supporting the hypotheses of:
  - local depletion in areas with longer exploitation histories?

Challenges

1. complex space-time data, with a spatially complicated domain
   - The spatial boundary of blue ling is complicated, depending on the topography of the sea bed.
   - We want to avoid smoothing accross boundary features, such as areas which are separated by a deep canyon.
   - Inappropriately imposing smoothness across boundary features might induce model mis-specification.
   - Will need to test whether space-time interactions are present.
2. fishery industry data - preferential sampling
3. continuous response with many zeros (19%)

Solutions

- Use a generalised additive mixed model (GAMM) incorporating a smooth function of space and time (1)
- tensor product from a soap film smooth of space and a penalised regression spline for time (1)
- model checking and validation to ensure there is no model mis-specification (1,2)
- Control for effects of fisheries management, targeting and species behaviour (2)
- Tweedie distribution (3)

Blue Ling model

- Reponse 'kg blue Ling in haul *i*', y_i, n = 17 614 observations
- Full model is

$$
\begin{aligned}
\log(\mu_i) = {} & f_1(\texttt{duration}_i) + f_2(\texttt{depth}_i, \texttt{year}_i) + v_{k(i)} \\
& + f_5(\texttt{depth}_i, \texttt{month}_i) + f_3(\texttt{depth}_i) + f_4(\texttt{month}_i) \\
& + f_6(\texttt{north}_i, \texttt{east}_i, \texttt{year}_i),
\end{aligned} \tag{1}
$$

- $\mu_i = \mathrm{E(y_i)}$ and $y_i \sim \mathrm{Tweedie}(\mu_i, \phi\mu_i^p)$, $p = 1.5$;
- $v_{k(i)}$ is a random vessel effect, assumed i.i.d. $N(0, \sigma_v^2)$;
- f_{1-6} are smooth functions of the covariates available with each haul.
- RED: effects of fisheries management and targetting;
- BLACK: biological effects.

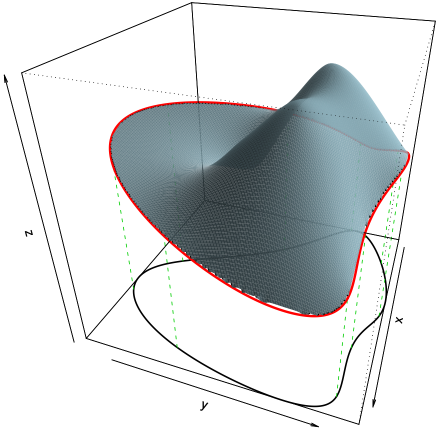
Blue Ling model ...cont'd

- $f_1(\texttt{duration}_i)$ and $f_3(\texttt{depth}_i)$: thin plate regression splines (TPRS);
- $f_4(\texttt{month}_i)$: cyclic cubic regression spline (CCRS)
- $f_2(\texttt{depth}_i, \texttt{year}_i)$: 2d tensor product of a TPRS and cubic regression spline (CRS);
- $f_5(\texttt{depth}_i, \texttt{month}_i)$: 2d tensor product of a tensor product of a TPRS and a CCRS.

$f_6(\mathtt{north}_i, \mathtt{east}_i, \mathtt{year}_i)$

- 3D tensor product of two *marginal* bases and penalties for time and space:
- $f_{n,e}(\mathtt{north}, \mathtt{east})$ and $f_y(\mathtt{year})$
  2-d isotropic smoother for space and a 1-d CRS for year
- allows spatial smooth to be isotropic while being invariant to relative scaling of space and time (Augustin et al, 2009).
- $f_{n,e}(\mathtt{north}, \mathtt{east})$
  - compare the performance of a **soap film smooth** with a **TPRS**
  - soap film smooth will respect the biological boundary, but require manual knot selection;
  - TPRS will smooth accross boundary features, possibly leading to model mis-specification, but no knot selection required.

Soap film smoother (Wood et al, 2008)



Tensor product smooths

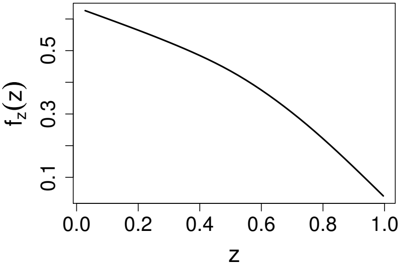
- A time varying spatial soap film can be constructed as a (pair of) tensor product smooth(s).
- Tensor product smooths are best explained using a 2D example.
- Consider constructing a smooth of x, z.
- Start by choosing *marginal* bases and penalties, as if constructing 1-D smooths of x and z. e.g.

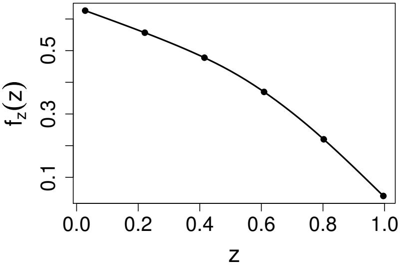
$$f_x(x) = \sum \alpha_i a_i(x), \quad f_z(z) = \sum \beta_j b_j(z),$$

$$J_x(f_x) = \int f_x''(x)^2 dx = \boldsymbol{\alpha}^\mathsf{T} \mathbf{S}_x \boldsymbol{\alpha} \ \ \& \ \ J_z(f_z) = \mathcal{B}^\mathsf{T} \mathbf{S}_z \mathcal{B}$$

Marginal reparameterization

- Suppose we start with $f_z(z) = \sum_{i=1}^{6} \beta_j b_j(z)$, on the left.

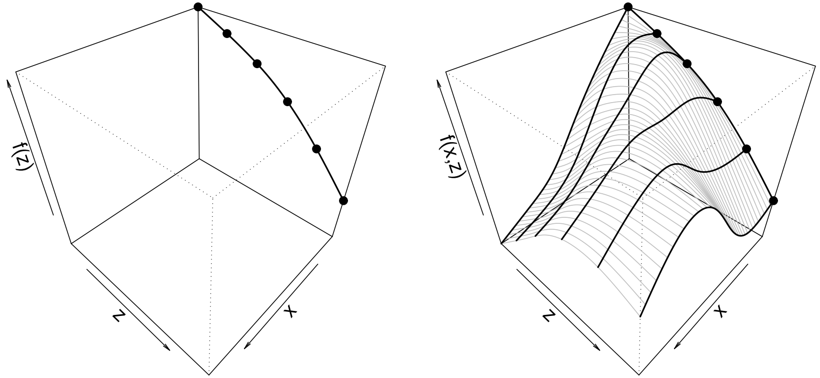




- We can always re-parameterize so that its coefficients are functions heights, at knots (right). Do same for f_x.

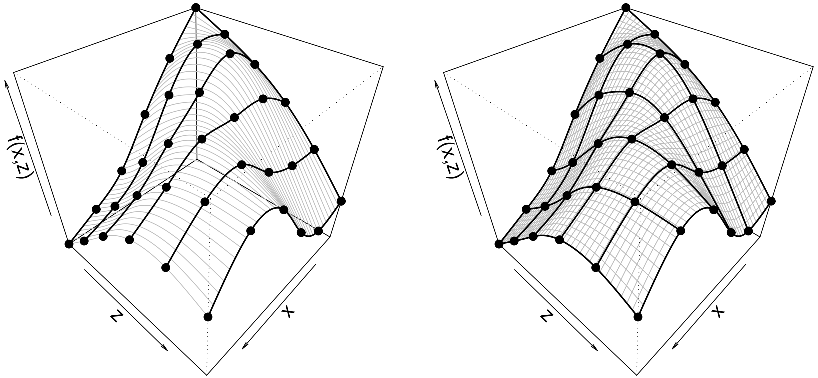
Making f_z depend on x

- Can make f_z a function of x by letting its coefficients vary smoothly with x



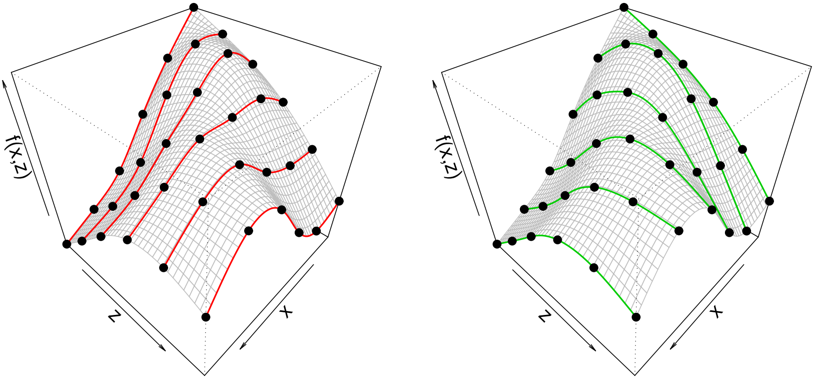
The complete tensor product smooth

- Use f_x basis to let f_z coefficients vary smoothly (left).
- Construct in symmetric (see right).



Tensor product penalties - one per margin

- x-wiggliness: sum marginal x penalties over red curves.
- z-wiggliness: sum marginal z penalties over green curves.



Tensor product expressions

- So the tensor product basis construction gives:

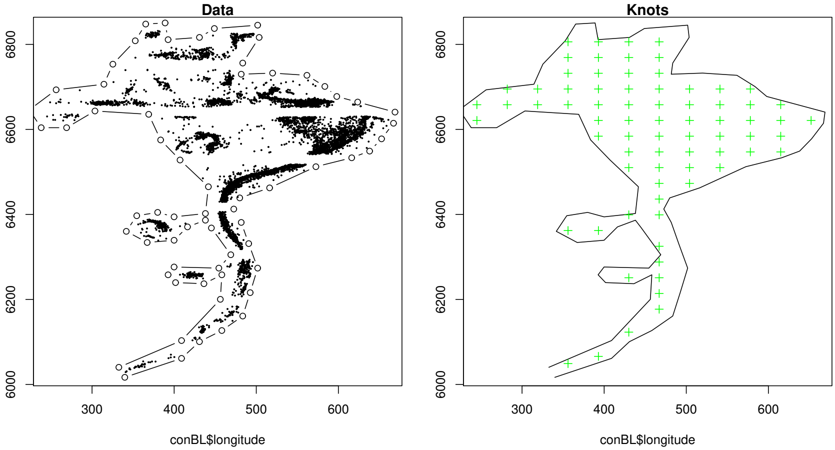
$$f(x,z) = \sum \sum \beta_{ij} b_j(z) a_i(x)$$

- With double penalties

$$J_z^*(f) = \boldsymbol{\beta}^{\mathsf{T}} \mathbf{I}_I \otimes \mathbf{S}_z \boldsymbol{\beta} \text{ and } J_x^*(f) = \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S}_x \otimes \mathbf{I}_J \boldsymbol{\beta}$$

- The construction generalizes to any number of marginals and multi-dimensional marginals.
- In particular a tensor product of a soap film and a 1D smooth of time is possible.
- The soap film smoother is separated into the boundary-interpolating-film and the deviation-from-film parts, and tensor products with time are formed for each.

Boundary and knots



Parameter estimation - use Bayesian representation of generalised linear mixed model (GLMM)

for hauls $y_i \sim \mathrm{EF}(\mu_i, \phi)$

$$
\begin{aligned}
g(\mu_i) &= f_1(\texttt{duration}_i) + f_2(\texttt{depth}_i, \texttt{year}_i) + v_{k(i)} \\
&\quad + f_5(\texttt{depth}_i, \texttt{month}_i) + f_3(\texttt{depth}_i) + f_4(\texttt{month}_i) \\
&\quad + f_6(\texttt{north}_i, \texttt{east}_i, \texttt{year}_i) \\
g(\mu_i) &= \mathbf{X}_i \boldsymbol{\beta}
\end{aligned}
$$

- $\mathbf{X}_i$ is a row of the model matrix of all model components of the model including all the basis functions evaluated at observations i ;
- parameter vector $\boldsymbol{\beta}$ contains all coefficients.

Parameter estimation - cont'd

- smoothing penality $\sum_j \lambda_j \boldsymbol{\beta}^\mathsf{T} \mathbf{S}_j \boldsymbol{\beta}$.
- (inproper) prior, each component is a Gaussian (intrinsic) random field

$$\boldsymbol{\beta} \sim N(\mathbf{0}, (\sum_j \lambda_j \mathbf{S}_j)^{-} \phi)$$

- estimates/posterior modes for $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmax}} \;\; \mathrm{l}(\boldsymbol{\beta}) - \frac{1}{2\phi} \sum_j \lambda_j \boldsymbol{\beta}^\mathsf{T} \mathbf{S}_j \boldsymbol{\beta}$$

  in practice use PIRLS
- posterior (large sample approximation)

$$\boldsymbol{\beta} | y \sim N(\hat{\boldsymbol{\beta}}, (\mathbf{X}^T \mathbf{W} \mathbf{X} + \sum_j \lambda_j \mathbf{S}_j)^{-1} \phi)$$

  with $\mathbf{W} = \text{diag}(w_i)$ - usual GLM weights

Parameter estimation - cont'd

- Find $\hat{\lambda}$ using a Laplace approximation of the Bayesian marginal log likelihood (REML, empirical Bayes)

$$\hat{\boldsymbol{\lambda}} = \underset{\lambda}{\text{argmax}} \log \int L(\beta) f(\beta) d\beta$$

  in practice use Newtons's method (with exact derivatives using implicit differentiation)
- nested iteration scheme is implemented in the gam() function of the **mgcv** R package.
- Estimate temporal trends with Bayesian credible intervals by sampling from the posterior of $\hat{\beta}$.

Model selection

Based on root mean square prediction error (RMSPE) estimated by cross-validation

1. Start with full model and check whether the random vessel effect can be replaced by a linear effect of vessel power.
2. Is space-time additive?
3. Is the soap smoother necessary for the spatial effect?

Models

Model (number)	Terms
Base	$f_1(duration_i) + f_2(depth_i, year_i)$ $+f_3(depth_i) + f_4(month_i) + f_5(depth_i, month_i)$
Full.ves (1a)	Base $+v_{k(i)} + f_6(north_i, east_i, year_i)$
Full.pow (2a)	Base $+\beta power_{k(i)} + f_6(north_i, east_i, year_i)$
Add.ves (3a)	Base $+v_{k(i)} + f_7(north_i, east_i) + f_8(year_i)$
Add.pow (4a)	Base $+\beta power_{k(i)} + f_7(north_i, east_i) + f_8(year_i)$
Nosoap.full.ves (1b)	Base $+v_{k(i)} + f_9(north_i, east_i, year_i)$
Nosoap.full.pow (2b)	Base $+\beta power_{k(i)} + f_9(north_i, east_i, year_i)$
Nosoap.add.ves (3b)	Base $+v_{k(i)} + f_{10}(north_i, east_i) + f_{11}(year_i)$
Nosoap.add.pow (4b)	Base $+\beta power_{k(i)} + f_{10}(north_i, east_i) + f_{11}(year_i)$

1a - 4a use soap smoother for space;
1b - 4b use thin-plate regression spline (TPRS) for space.

Model statistics and 11-fold cross validation results.

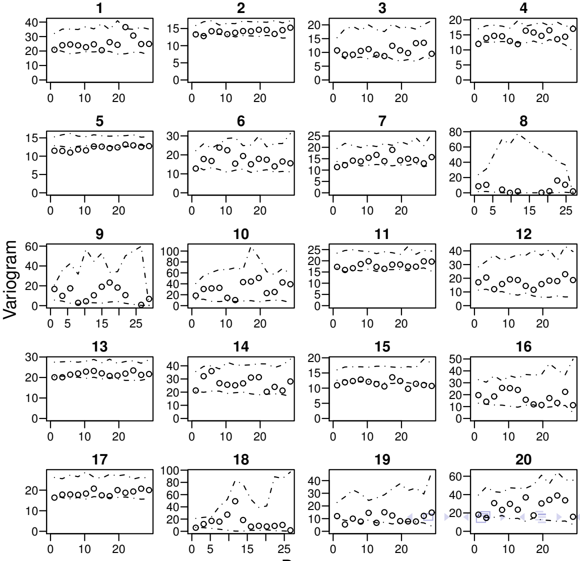
Model (number)	edf	$RMSPE^R$ overall	$RMSPE^V$ overall	$RMSPE^R$ edge6	new5	new6	other6	ref5
Full.ves (1a)	335.98	699.43	705.08	431.73	1115.22	**843.57**	445.87	1065.99
Full.pow (2a)	319.15	**695.80**	**702.42**	436.08	1078.32	849.82	446.02	**1063.30**
Add.ves (3a)	208.95	711.04	705.13	427.55	1080.71	875.78	426.98	1121.14
Add.pow (4a)	210.17	708.33	702.78	430.64	**1062.11**	883.37	426.82	1115.93
Nosoap.full.ves (1b)	376.23	700.13	705.28	425.99	1132.29	855.42	431.43	1073.31
Nosoap.full.pow (2b)	362.44	697.49	704.74	430.17	1100.65	860.24	431.53	1071.76
Nosoap.add.ves (3b)	218.68	707.61	704.78	**424.50**	1087.72	880.94	423.22	1111.81
Nosoap.add.pow (4b)	202.19	705.45	703.71	427.74	1074.07	886.84	**423.05**	1106.74

Selected model: Full.pow (2a)

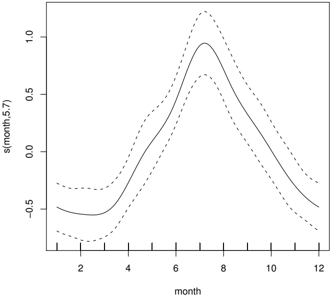
Selected model

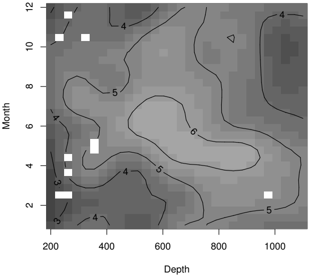
$$\log(\mu_i) = f_1(\texttt{duration}_i) + f_2(\texttt{depth}_i, \texttt{year}_i) + \beta \textit{power}_k$$
$$f_3(\texttt{depth}_i) + f_4(\texttt{month}_i) + f_5(\texttt{depth}_i, \texttt{month}_i)$$
$$+ f_6(\texttt{north}_i, \texttt{east}_i, \texttt{year}_i), \quad (2)$$

Model checking - Empirical variograms of Pearson residuals per vessel along the time-axis.

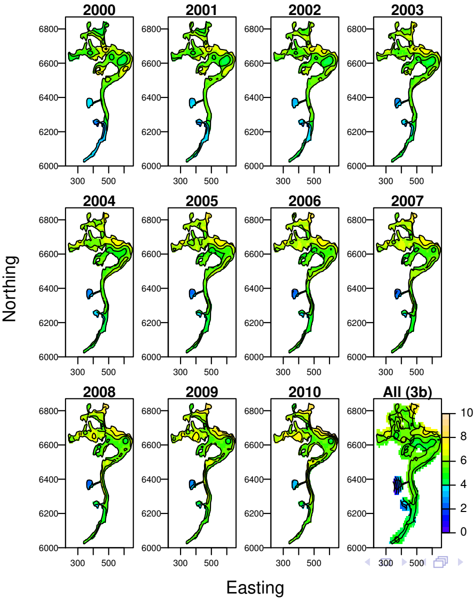


Model effects: Month smooth with ± two standard errors (left); Depth-month smooth (right).

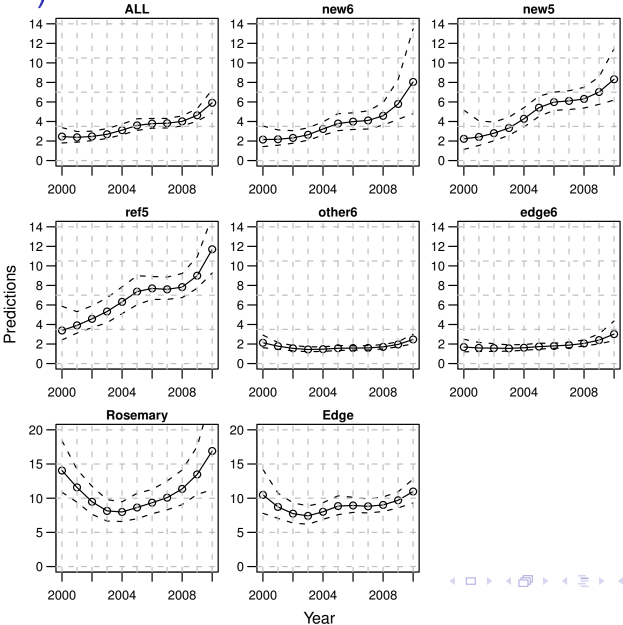




Fitted spatial model smooths by year for model 2a (soap smoother) and model 3b (TPRS smoother)



Time trends by area (January) and for two spawning areas (April)



Some conclusions

- 3d tensor product of soap film smooth for space and CRS smooth for year allowed us to test for the presence of space-time interaction.
- Soap film smoother ensures that we are not smoothing accross the complicated boundary.
- What is the time trend of blue ling abundance? It appears to be constant/increasing.
- Is there any evidence for localised depletions in areas with a longer exploitation history? yes - maybe. Space-time interaction term is required.
- Cannot assess how biased our results since data are based on preferential sampling. But our model allows to control for effects of fisheries management and targetting.

Kai et al (2017) - Spatio-temporal distribution of pelagic sharks

Model catch y_i of set i with $\mu_i = \mathrm{E}(y_i)$ with $y_i \sim \mathrm{EF}(\mu_i, \phi)$

$$\mu_i = \mathrm{effort}_i \;\; d(s_i, t_i, q_i)$$
$$\log(\mu_i) = \log(\mathrm{effort}_i) + \alpha_{1t_i} + \alpha_{2q_i} + \gamma(s_i) + \theta(s_i, t_i) + \omega(s_i, q_i) + \beta_1 \mathrm{SST}_j + \beta_2 \mathrm{SST}_j^2$$

- with station s_i (at 1^o lat lon resolution), t_i=1,...,20 (year quarters of 2010 - 2015), quarter q_i;
- α_{1t_i}, α_{2q_i} are factor variables for the year quarters of 2010 to 2015 and quarter respectively;
- $\gamma \sim \mathrm{MVN}(0, \sigma_\gamma^2 \mathrm{R}_{spatial})$ GMRF, i.e. a random effect for space with Matérn correlation structure;
- $\theta(s_i, t_i) \sim \mathrm{MVN}(0, \sigma_\theta^2 \mathrm{R}_{spatial}^\kappa \otimes \mathrm{R}_{\mathrm{AR1}}^\rho)$,
- $\omega(s_i, q_i) \sim \mathrm{MVN}(0, \sigma_\omega^2 \mathrm{R}_{spatial}^\kappa \otimes \mathrm{R}_{\mathrm{AR1}}^\rho)$,

Connection between different approaches

- Smooths, random effects and Gaussian (Markov) random fields are equivalent (Kimeldorf and Wahba, 1970, Silverman, 1985).
- Mixed model software can be used to estimate smooths/GAMs and conversely software for estimating smooths can be used to estimate Gaussian random effects (Verbyla et al, 1999; Ruppert et al, 2003);
- Smoother penalty matrix is equivalent to assumed precision matrix of the MVN prior of the random field;

Connection/comparison cont'd

- **TMB** and **mgcv** use first order Laplace approximation to marginal likelihood; **INLA** uses higher order Laplace approximation
  - if using a GMRF smoother **TMB** and **INLA** exploit sparsity, **mgcv** uses a reduced rank approx. of it.
- **mgcv** - computational efficiency through reduced rank method.
  Efficient with big data, many predictors with non-linear effects and a standard model structure - can only use built in distributions.
- **INLA** - efficient because of sparsity of precision matrix; can only use the built in distributions and smoothers (Gaussian Markov random fields);
- **TMB** - very flexible, efficient for non-standard models; but high memory footprint (backward automatic differentiation).

References

Augustin, N.H., Trenkel, V., Wood, S.N. and Lorance, P. Space-time modelling of blue ling for fisheries stock management. *Environmetrics* (accepted).

Augustin, N.H. and Musio, M. and von Wilpert, K. and Kublin, E. and Wood, S.N. and Schumacher, M. 2009. Modelling spatio-temporal forest health monitoring data. *Journal of the American Statistical Association*, 104 (487), 899-911.

Kai et al

GS Kimeldorf, G Wahba. 1970. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. The Annals of Mathematical Statistics 41 (2), 495-502.

Silverman, B.W., 1985. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. Journal of the Royal Statistical Society. Series B (Methodological), pp.1-52.

Tweedie, M. C. K., 1984. An index which distinguishes between some important exponential families. Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference, 579-604,editors J. K. Ghosh and J. Roy, Calcutta: Indian Statistical Institute.

Verbyla, A.P., Cullis, B.R., Kenward, M.G. and Welham, S.J., 1999. The analysis of designed experiments and longitudinal data by using smoothing splines. Journal of the Royal Statistical Society: Series C (Applied Statistics), 48(3), 269-311.

Wood, S.N. and Bravington, M.V. and Hedley, S.L., 2008. Soap film smoothing. *Journal of the Royal Statistical Society*,70 (5), 931-955.

Wood, S.N. 2017. Generalized additive models. An Introduction with R. CRC Press, Boca Raton.