Predicting loggerhead turtle abundance in the North Western Mediterranean Sea

Designing Robust Species Distribution Models

Matthieu Authier

Observatoire PELAGIS UMS-CNRS 3462

14 December 2017

L'Observatoire PELAGIS

Since 2011

Observatoire PELAGIS
Unité Mixte de Service UMS 3462, Université de La Rochelle & CNRS

Since 2011

Observatoire PELAGIS
Unité Mixte de Service UMS 3462, Université de La Rochelle & CNRS

1. Observatoire
2. Expertise

Since 2011

Observatoire PELAGIS
Unité Mixte de Service UMS 3462, Université de La Rochelle & CNRS

1. Observatoire
2. Expertise
3. Recherche → SEC-LR (UMR 7372, 10 chercheurs)

Actions

Actions
principale

spécifiques

Actions

Actions	Observatoire
principale	Échouages
spécifiques	MEGASCOPE SAMM & REMMOA Dunkrisk

Actions

Actions	Observatoire	Expertise
principale	Échouages	CBI, CMR, ...
spécifiques	MEGASCOPE SAMM & REMMOA Dunkrisk	DCSMM DHFF EMR

Actions

Actions	Observatoire	Expertise
principale	Échouages	CBI, CMR, ...
spécifiques	MEGASCOPE SAMM & REMMOA Dunkrisk	DCSMM DHFF EMR

DATA

Acquisition, Nettoyage, Validation, Stockage, Analyses, Diffusion...

SAMM & REMMOA

Observatoire: Campagnes Aériennes

Britten Norman 2 affrété pour la campagne
(G.Dorémus - AAMP/Observatoire PELAGIS)

Observateur positionné dans le hublot-bulle
(T. Auger - AAMP/Observatoire PELAGIS)

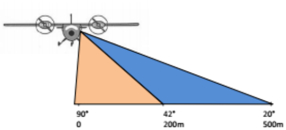
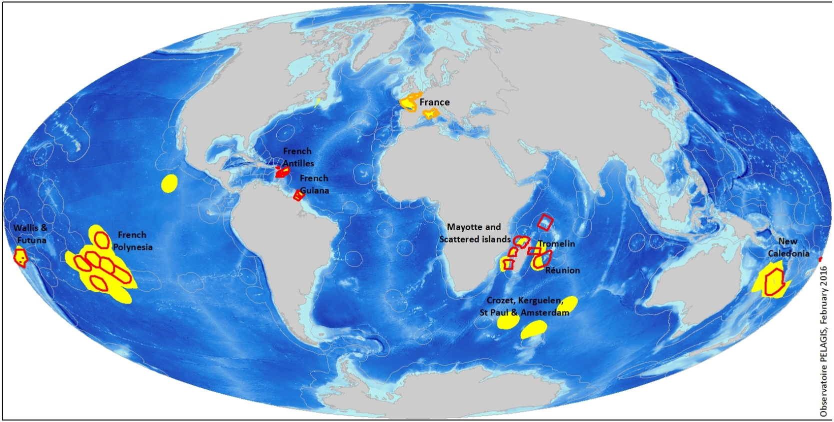


Figure 1. Angles d'observation et distances correspondantes à partir des hublots-bulles.

Dauphins de Risso vu d'avion
(M. Perri - AAMP/Observatoire PELAGIS)

Observatoire: Campagnes Aériennes



Spatial Planning

’Torremolinos Charter’ adopted in 1983 by the European Conference of Ministers responsible for Regional Planning

"Spatial planning gives geographical expression to the economic, social, cultural and ecological policies of society. It is a scientific discipline, an administrative technique and a policy developed as an interdisciplinary and comprehensive approach directed towards a balanced regional development and the physical organisation of space according to an overall strategy."

Spatial Planning

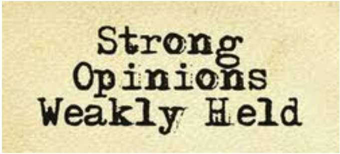
’Torremolinos Charter’ adopted in 1983 by the European Conference of Ministers responsible for Regional Planning

"Spatial planning gives geographical expression to the economic, social, cultural and ecological policies of society. It is a scientific discipline, an administrative technique and a policy developed as an interdisciplinary and comprehensive approach directed towards a balanced regional development and the physical organisation of space according to an overall strategy."

$\Rightarrow$ crucial for biodiversity conservation and policies (*e.g.* MSFD, . . .)

Species Distribution Models

Disclaimer



SDM

Predict from environmental inputs ($x_{k_{\in[1:p]}}$) where a species of interest occurs

$$\mathbb{E}[\text{Response Variable}] = f(\text{Environmental Inputs})$$

SDM

Predict from environmental inputs ($x_{k_{\in[1:p]}}$) where a species of interest occurs

$\mathbb{E}$[Response Variable] = f(Environmental Inputs)

A SDM is typically a mathematical statement ("specification") about the Conditional Expectation Function CEF:

SDM

Predict from environmental inputs ($x_{k \in [1:p]}$) where a species of interest occurs

$\mathbb{E}$[Response Variable] = f(Environmental Inputs)

A SDM is typically a mathematical statement ("specification") about the Conditional Expectation Function CEF:

1. linear reg. CEF: $\mathbb{E}[Y|X] = \beta_0 + \sum_{k=1}^{p} \beta_k \times x_k$
2. logistic reg. CEF: $\mathbb{E}[Y|X] = \dfrac{1}{1 + e^{-\beta_0 - \sum_{k=1}^{p} \beta_k \times x_k}}$
3. linear gam. CEF: $\mathbb{E}[Y|X] = \beta_0 + \sum_{k=1}^{p} s_k(x_k)$
4. *etc...*

SDM

Analytical workflow

Observed Data time$_t$ (Long, Lat)$_{\mathrm{obs}}$	$\longrightarrow$	Modelling	$\longrightarrow$	Predictions
Occurrence Habitat use Abundance Inputs $(x_1, .., x_p)$	$\searrow$ $\nearrow$	CEF	$\nearrow$ $\searrow$	Spatial (Long, Lat)$_{\mathrm{pred}}$ Temporal time$_{\mathbf{t+1}}$ (Péron et al., 2012)

SDM: Usual Study Design

1. collect dataset Y of size n
2. extract p environmental covariates at sample locations: X

SDM: Usual Study Design

1. collect dataset Y of size n
2. extract p environmental covariates at sample locations: X
3. exclude combination of covariates with pairwise correlation $>$ some threshold (e.g. 0.7)
4. specification search: choose CEF minimizing *e.g.* AIC $\propto \log \ell(Y|\hat{\theta})$

SDM: Usual Study Design

1. collect dataset Y of size n
2. extract p environmental covariates at sample locations: X
3. exclude combination of covariates with pairwise correlation $>$ some threshold (e.g. 0.7)
4. specification search: choose CEF minimizing *e.g.* $\text{AIC} \propto \log \ell(Y|\hat{\theta})$
5. check fit and predictive accuracy

SDM: Usual Study Design

1. collect dataset Y of size n
2. extract p environmental covariates at sample locations: X
3. exclude combination of covariates with pairwise correlation $>$ some threshold (e.g. 0.7)
4. specification search: choose CEF minimizing *e.g.* $\text{AIC} \propto \log \ell(Y|\hat{\theta})$
5. check fit and predictive accuracy
6. predict from selected model (or model sets) and X_{new}

Predicting from SDM

More often than not, interest lies in predictions in **unsampled** locations

Predicting from SDM

More often than not, interest lies in predictions in **unsampled** locations

In usual framework, robustness is hoped for after checking the specification search:

- checking is internal (*e.g.* use in-sample cross validation to estimate out-of-sample predictive accuracy);
- ignores data collection design (random partitioning of the data).

Predicting from SDM

More often than not, interest lies in predictions in **unsampled** locations

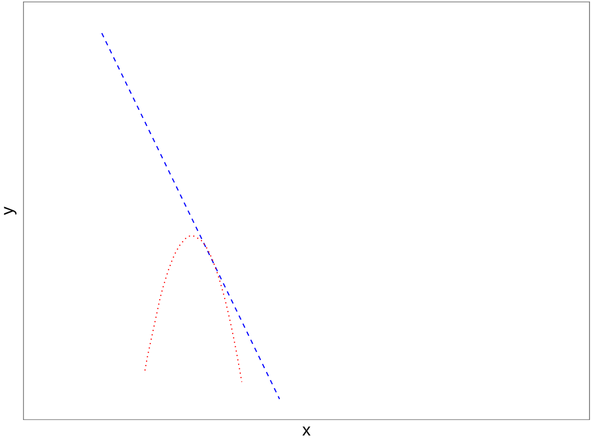
In usual framework, robustness is hoped for after checking the specification search:

- checking is internal (*e.g.* use in-sample cross validation to estimate out-of-sample predictive accuracy);
- ignores data collection design (random partitioning of the data).

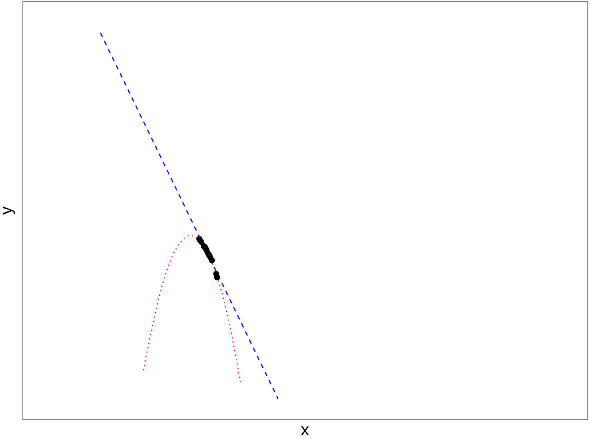
Fundamental problem:
Predictions can be heavily model-dependent, that is **non-robust**.

Prediction

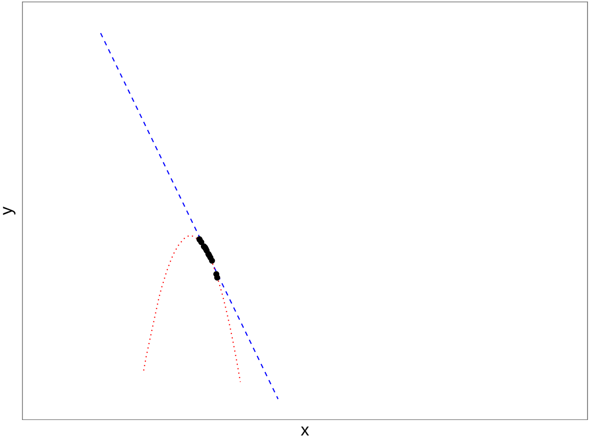
Specification Search



Specification Search

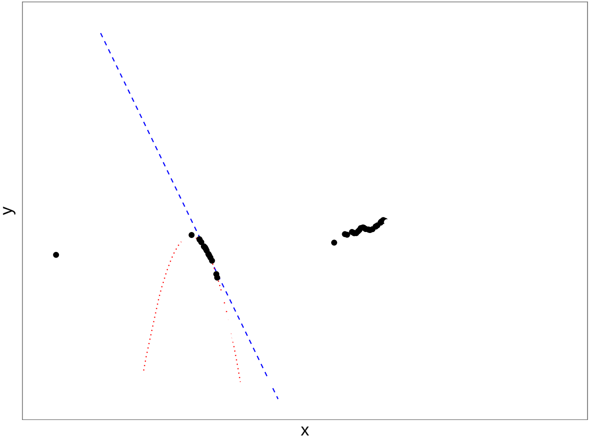


Specification Search

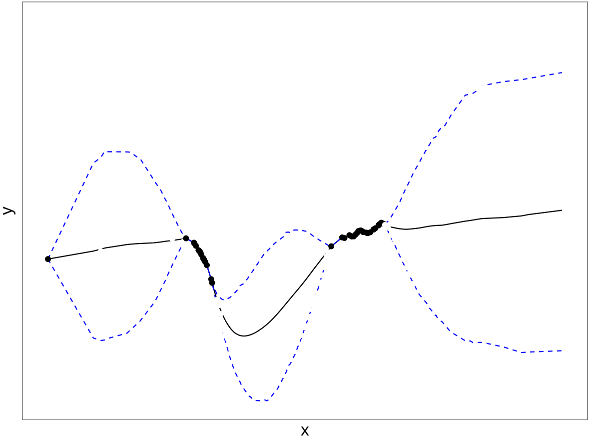


Both $R^2 \approx .0.99$, yet very different predictions...

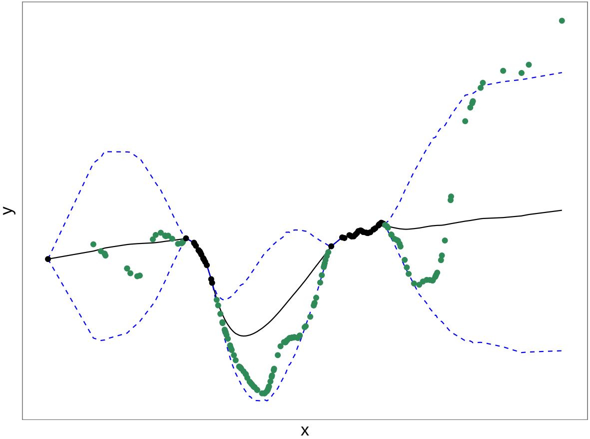
Extrapolations



Interpolations and Extrapolations



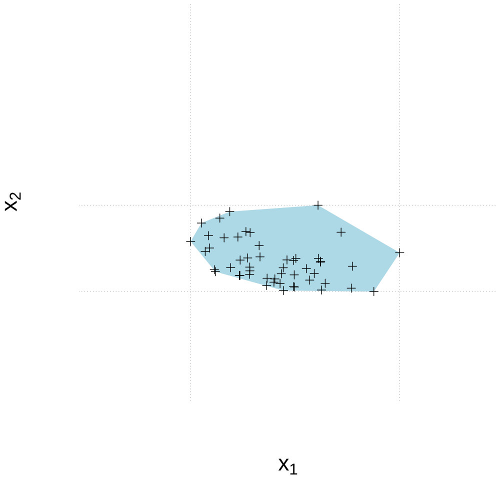
Interpolations and Extrapolations

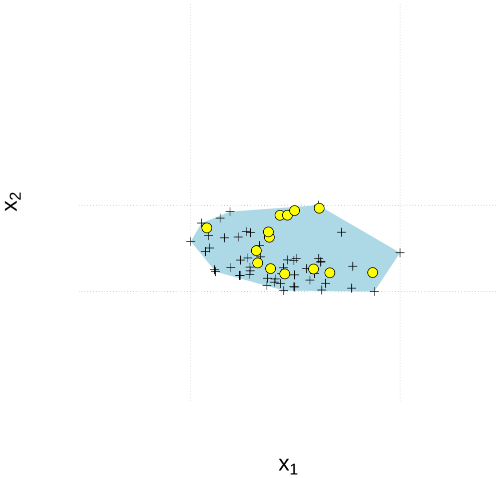


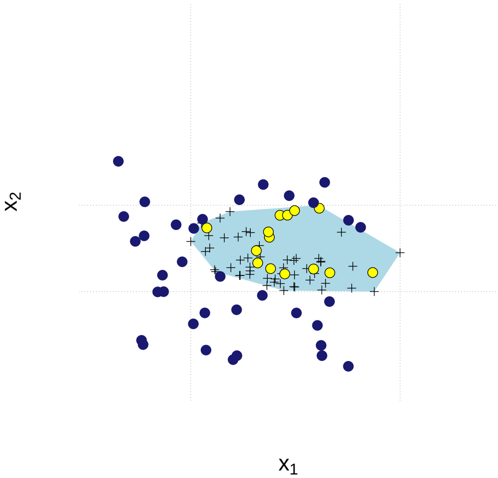
Interpolations and Extrapolations

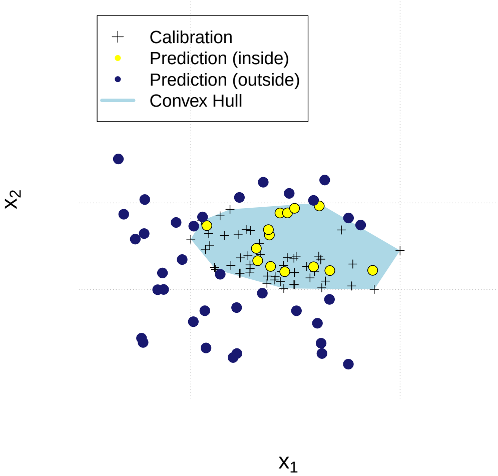
How can we know
what kind of predictions
we are making?

Convex Hull


x2
x1


x_2
x_1


X_2
X_1


Calibration
Prediction (inside)
Prediction (outside)
Convex Hull
X_2
X_1

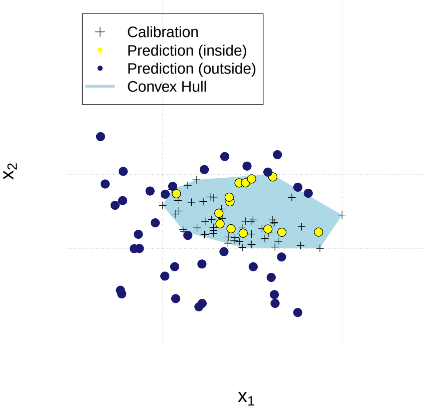
Gower's Distance

Gower's nonparametric measure $G^2_{i,j}$

average absolute distance between i and j, divided by the range $r_k = \max(x_{.k}) - \min(x_{.k})$

$$G^2_{i,j} = \frac{1}{K} \times \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{r_k} \qquad (1)$$

(King & Zeng, 2007)



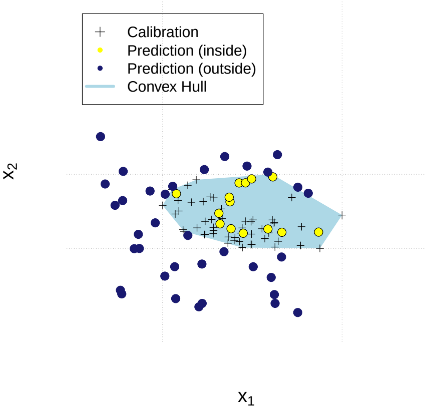
Gower's nonparametric measure $G^2_{i,j}$

average absolute distance between i and j, divided by the range $r_k = \max(\mathrm{x}_{.\mathrm{k}}) - \min(\mathrm{x}_{.\mathrm{k}})$

$$G^2_{i,j} = \frac{1}{K} \times \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{r_k} \qquad (1)$$

(King & Zeng, 2007)

No Need of Y!



Convex Hull Computations

R package `WhatIf` (Stoll et al., 2009)

```
whatif( formula    = NULL,
        data       = calibrationData,
        cfact      = predictionData
        )
```

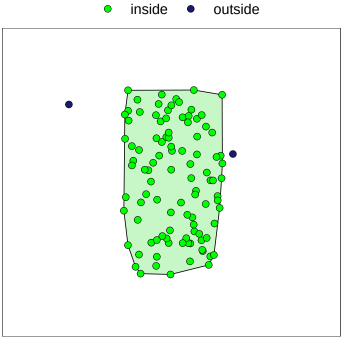
Convex Hull Computations

R package `WhatIf` (Stoll et al., 2009)

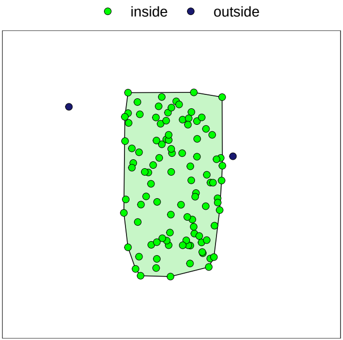
```
whatif(  formula  = NULL,
         data     = calibrationData,
         cfact    = predictionData
         )
```

Works with *x* continuous, categorical or both

Trust Thy Neighbours

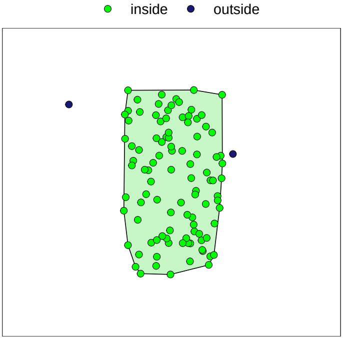


Trust Thy Neighbours



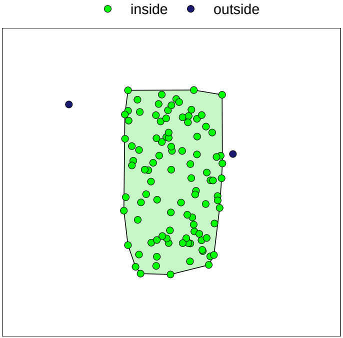
- G^2 = distance between two points as a proportion of the data range

Trust Thy Neighbours



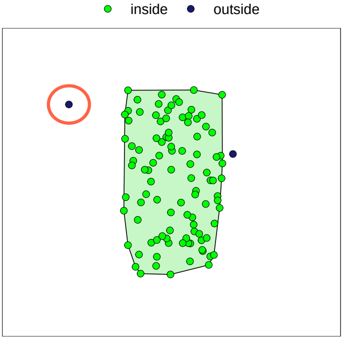
- G^2 = distance between two points as a proportion of the data range
- $G^2 = 0.3$ means the two points are 30% of the range apart

Trust Thy Neighbours



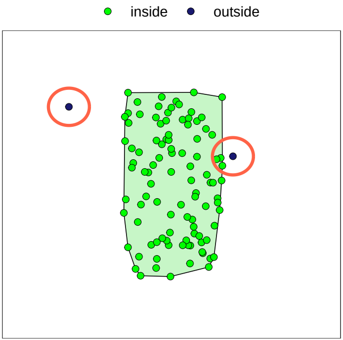
- G^2 = distance between two points as a proportion of the data range
- $G^2 = 0.3$ means the two points are 30% of the range apart
- can define a neighbourhood as the % points within a given radius ($\bar{G}^2$)

Trust Thy Neighbours



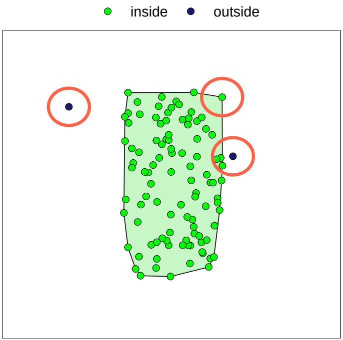
- G^2 = distance between two points as a proportion of the data range
- $G^2 = 0.3$ means the two points are 30% of the range apart
- can define a neighbourhood as the % points within a given radius ($\bar{G}^2$)

Trust Thy Neighbours



- G^2 = distance between two points as a proportion of the data range
- $G^2 = 0.3$ means the two points are 30% of the range apart
- can define a neighbourhood as the % points within a given radius ($\bar{G}^2$)

Trust Thy Neighbours



- G^2 = distance between two points as a proportion of the data range
- $G^2 = 0.3$ means the two points are 30% of the range apart
- can define a neighbourhood as the % points within a given radius ($\bar{G}^2$)

Convex Hull Computations

Assess with Gower's distance
for a prediction with X_{new}
(1) whether it is an interpolation or an extrapolation wrt X,
(2) how many neighbours in X it has; **without** doing any actual model fitting!

SDM: Alternate Study Design

1. collect dataset Y of size n
2. extract p environmental covariates at sampled locations: X
3. exclude combination of covariates with pairwise correlation $>$ some threshold (e.g. 0.7)

SDM: Alternate Study Design

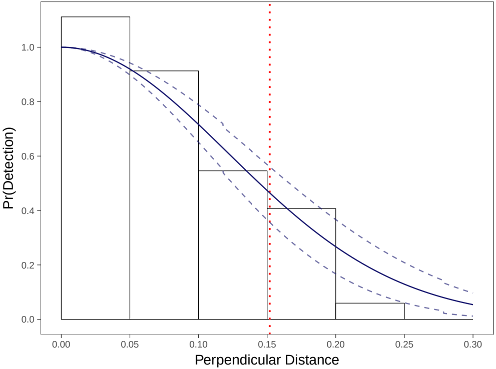
1. collect dataset Y of size n
2. extract p environmental covariates at sampled locations: X
3. exclude combination of covariates with pairwise correlation > some threshold (e.g. 0.7)
4. specification search: choose CEF minimizing extrapolation from X to X_{new}

SDM: Alternate Study Design

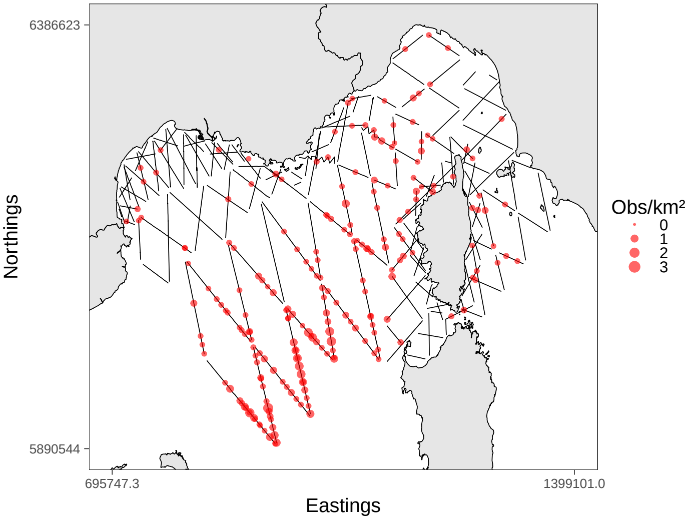
1. collect dataset Y of size n
2. extract p environmental covariates at sampled locations: X
3. exclude combination of covariates with pairwise correlation > some threshold (e.g. 0.7)
4. specification search: choose CEF minimizing extrapolation from X to X_{new}
5. check fit and predictive accuracy
6. predict from selected model at new locations X_{new}

Case Study: Loggerheads in the Mediterranean Sea

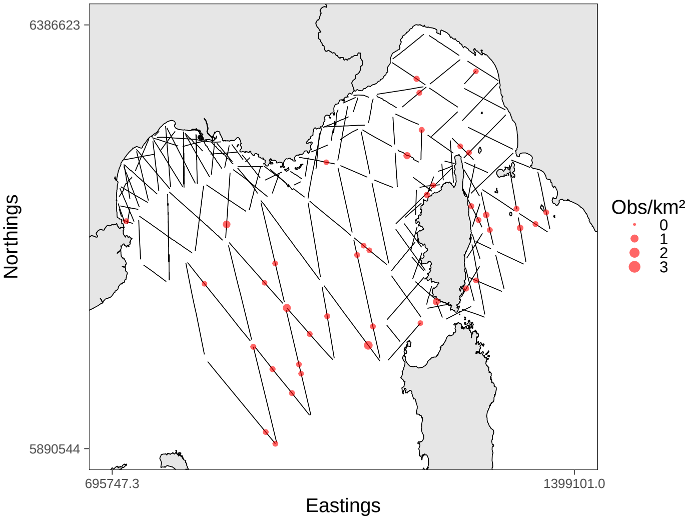
Loggerhead turtles



Summer survey: 308 obs



Winter survey: 49 obs



SDM

Search for a specification with 4 environmental covariates among 10 candidates

210 specifications, 101 with max. pairwise correlation < 0.7

SDM

Search for a specification with 4 environmental covariates among 10 candidates

210 specifications, 101 with max. pairwise correlation < 0.7

$$y_i \sim \mathcal{ZIP}(\alpha_i, \mathrm{Effort}_i \times e^{\mu_i})$$

SDM

Search for a specification with 4 environmental covariates among 10 candidates

210 specifications, 101 with max. pairwise correlation < 0.7

$$y_i \sim \mathcal{ZIP}(\alpha_i, \mathrm{Effort}_i \times e^{\mu_i})$$

$$\alpha_i = \mathrm{logit}^{-1}(\gamma_0 + \gamma_1 \times \mathrm{linear\ effort}_i + \gamma_2 \times \mathrm{Beaufort}_i)$$

SDM

Search for a specification with 4 environmental covariates among 10 candidates

210 specifications, 101 with max. pairwise correlation < 0.7

$$y_i \sim \mathcal{ZIP}(\alpha_i, \mathrm{Effort}_i \times e^{\mu_i})$$

$$\alpha_i = \mathrm{logit}^{-1}(\gamma_0 + \gamma_1 \times \mathrm{linear\ effort}_i + \gamma_2 \times \mathrm{Beaufort}_i)$$

$$\mu_i = \beta_0 + \sum_{k=1}^{4} \mathrm{BS}_k(x_{ik})$$

where $\mathrm{BS}_k(.)$ are cubic Bézier-splines (Eilers & Marx, 2010) with 10 knots.

SDM

Search for a specification with 4 environmental covariates

SDM

Search for a specification with 4 environmental covariates

Model fitting with `Stan` (Carpenter et al., 2017)

SDM

Search for a specification with 4 environmental covariates

Model fitting with Stan (Carpenter et al., 2017)

weakly informative normal priors with non-centered parametrization

SDM

Search for a specification with 4 environmental covariates

Model fitting with Stan (Carpenter et al., 2017)

weakly informative normal priors with non-centered parametrization

Gamma-Gamma mixture priors (Griffin & Brown, 2016) for variances:

$$\sigma^2|\lambda, \phi \sim \Gamma(\lambda, \phi)$$
$$\phi|\rho, s^2 \sim \Gamma(\rho, s^2)$$

SDM

Search for a specification with 4 environmental covariates

Model fitting with Stan (Carpenter et al., 2017)

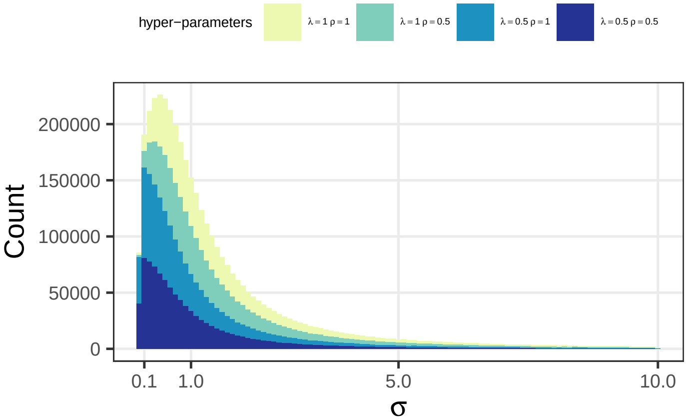
weakly informative normal priors with non-centered parametrization

Gamma-Gamma mixture priors (Griffin & Brown, 2016) for variances:

$$\sigma^2|\lambda, \phi \sim \Gamma(\lambda, \phi)$$
$$\phi|\rho, s^2 \sim \Gamma(\rho, s^2)$$

With $\lambda = 0.5$ and $\rho = 1.0$,
this prior has a mean of s^2 and a spike at 0.

Gamma-Gamma mixture priors



SDM

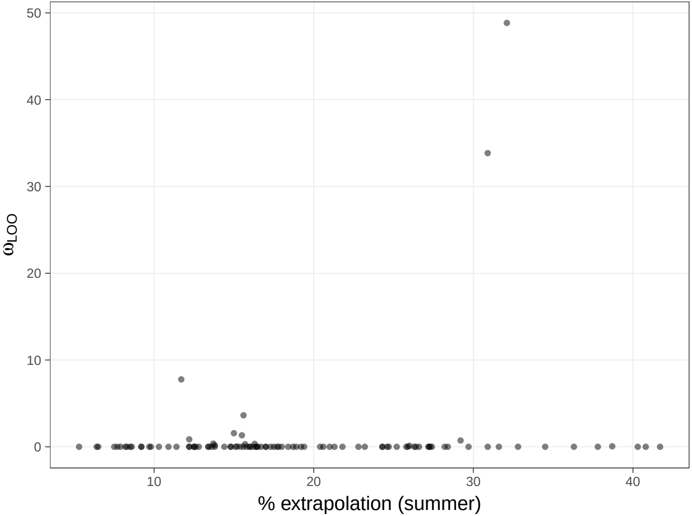
Search for a specification with 4 environmental covariates

SDM

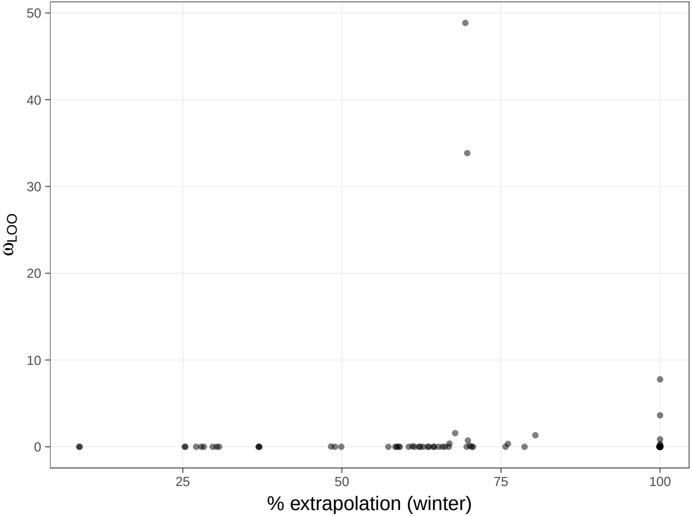
Search for a specification with 4 environmental covariates

1. LOO (Vehtari et al., 2017): Bathymetry, Distance to Shelf Break, NPP, Sea Level Anomaly

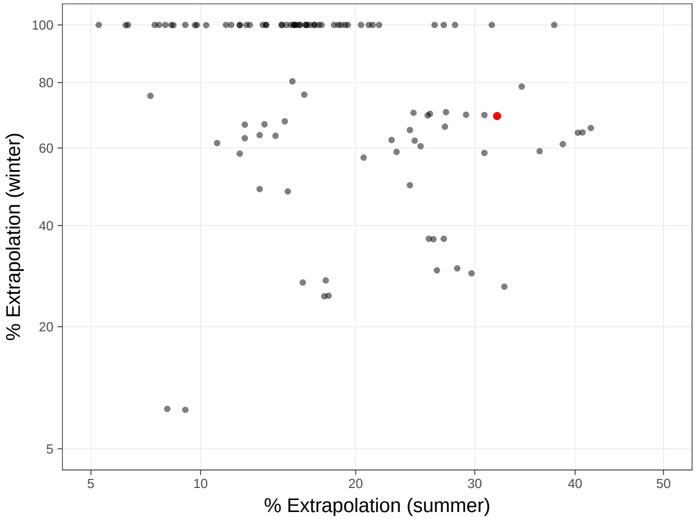
Extrapolation in Summer



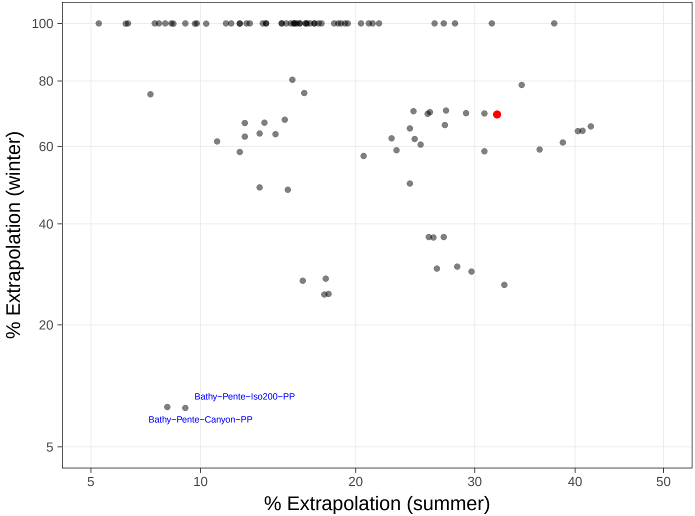
Extrapolation in Winter



Extrapolation



Extrapolation



SDM selection

Specification search for SDM with 4 environmental covariates

1. LOO: **Bathymetry**, **Distance to Shelf Break**, **NPP**, Sea Level Anomaly
2. Gower: **Bathymetry**, Slope, **Distance to Shelf Break**, **NPP**

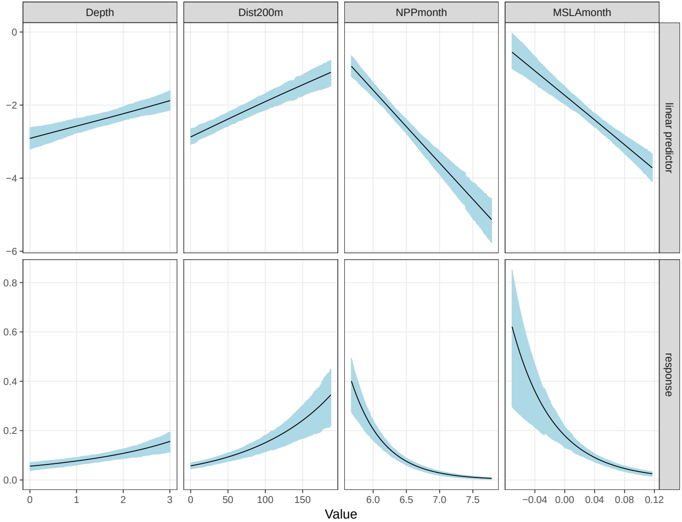
SDM selection

Specification search for SDM with 4 environmental covariates

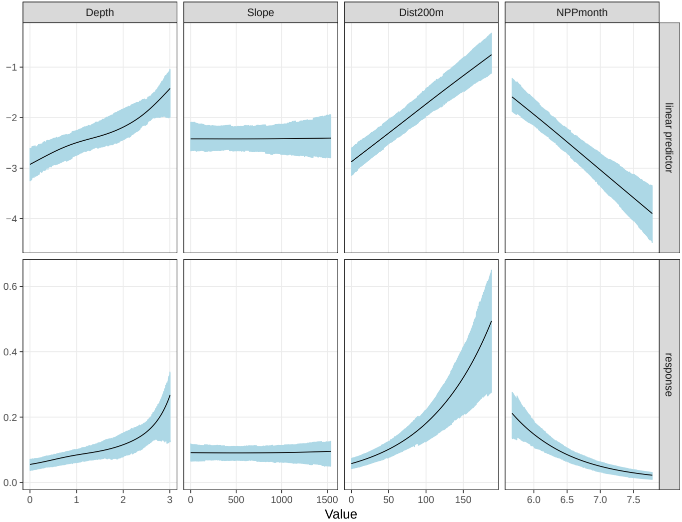
1. LOO: **Bathymetry**, **Distance to Shelf Break**, **NPP**, Sea Level Anomaly
2. Gower: **Bathymetry**, Slope, **Distance to Shelf Break**, **NPP**

1. $\hat{\omega}_{\text{LOO}} = 0.49$
2. $\hat{\omega}_{\text{LOO}} \approx 5 \times 10^{-7}$

Covariate Effects I



Covariate Effects II



Validation

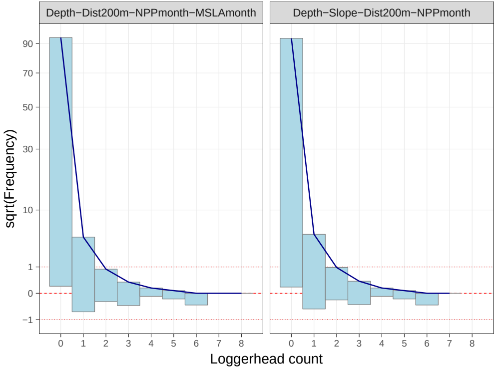
→ Two quantitative criteria: RMSE and Interval Score $\mathrm{INT}_\alpha(l, u, y_{\mathrm{pred}})$

1. $\mathrm{RMSE} = \sqrt{\sum_i \left((y_{\mathrm{obs}} - y_{\mathrm{pred}})^2\right)}$
2. $\mathrm{INT}_\alpha(l, u, y_{\mathrm{pred}}) = u - l + \frac{\alpha}{2}(l - y_{\mathrm{pred}})\mathbf{1}\{y_{\mathrm{pred}} < l\} + \frac{\alpha}{2}(y_{\mathrm{pred}} - u)\mathbf{1}\{y_{\mathrm{pred}} > u\}$

→ One graphical check: rootograms (Kleiber & Zeileis, 2016)

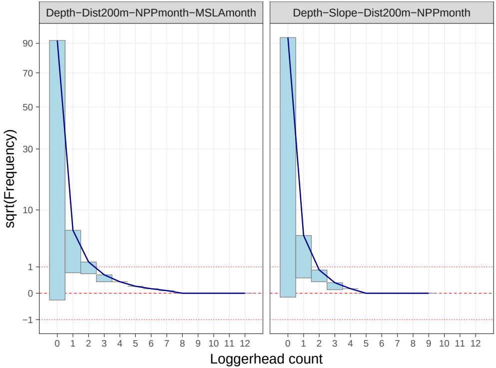
In-Sample GOF

RMSE	${}_{0.564}0.586_{0.607}$	${}_{0.570}0.594_{0.616}$
INT_α	809.0	809.5

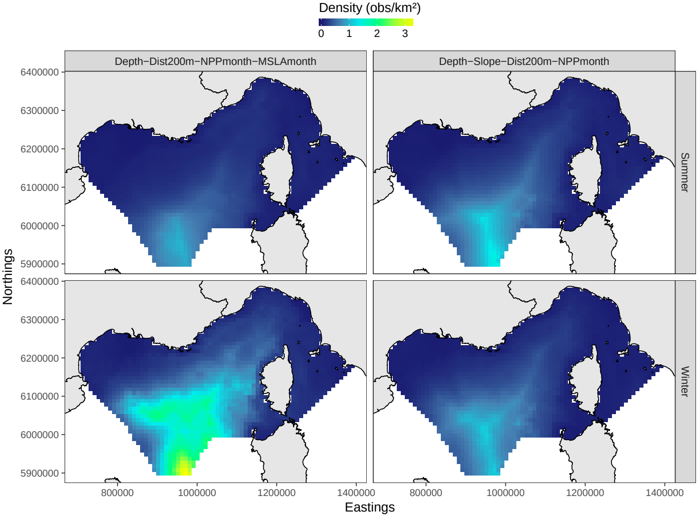


Out-of-Sample validation

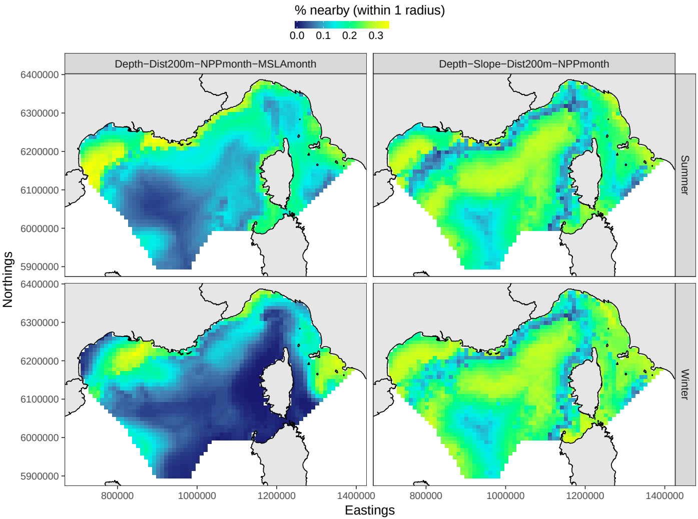
RMSE	$_{0.387}0.493_{0.587}$	$_{0.314}0.360_{0.408}$
INT_{α}	271.5	159.0



Predictions



Interpolations



Another Way?

SDM: tweaking the likelihood

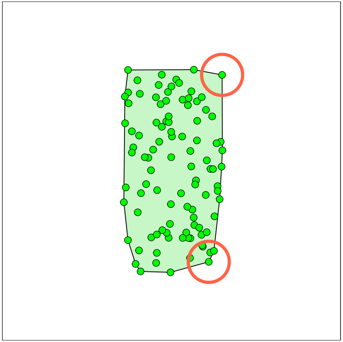
1. collect dataset Y of size n
2. extract p environmental covariates at sampled locations: X
3. exclude combination of covariates with pairwise correlation > some threshold (e.g. 0.7)

5. check fit and predictive accuracy
6. predict from selected model at new locations X_{new}

SDM: tweaking the likelihood

1. collect dataset Y of size n
2. extract p environmental covariates at sampled locations: X
3. exclude combination of covariates with pairwise correlation > some threshold (e.g. 0.7)
4. specification search:
   4.1 estimate a "neighbourhood" w_i of X
   4.2 use a weighted likelihood framework $\ell(Y|\hat{\theta})^w$
5. check fit and predictive accuracy
6. predict from selected model at new locations X_{new}

SDM with weighted likelihood



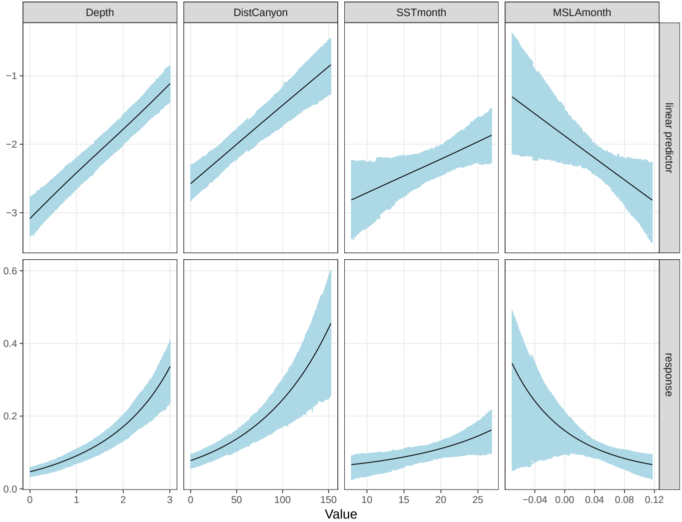
- G^2 = distance between two points as a proportion of the data range
- a neighbourhood is the % points within a one $\bar{G}^2$ radius
- $w_i = \dfrac{\text{size of neighbourhood}}{\text{average neighbourhood}}$ so that $n = \sum_i w_i$

SDM

Specification search for SDM with 4 environmental covariates

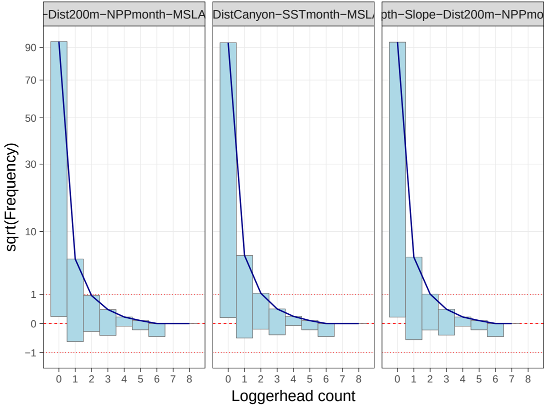
1. ℓ: **Bathymetry**, Distance to Shelf Break, NPP, Sea Level Anomaly
2. ℓ^w: **Bathymetry**, Distance to Canyon, SST, Sea Level Anomaly
3. Gower: **Bathymetry**, Slope, Distance to Shelf Break, NPP

Covariate Effects III



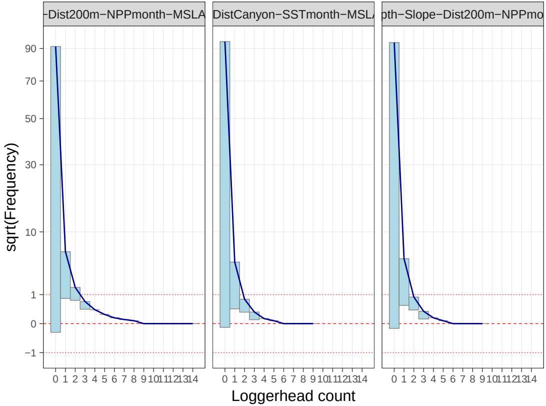
In-Sample GOF

RMSE	${}_{0.564}0.586_{0.607}$	${}_{0.566}0.601_{0.626}$	${}_{0.570}0.594_{0.616}$
INT_{α}	809.0	811.0	809.5

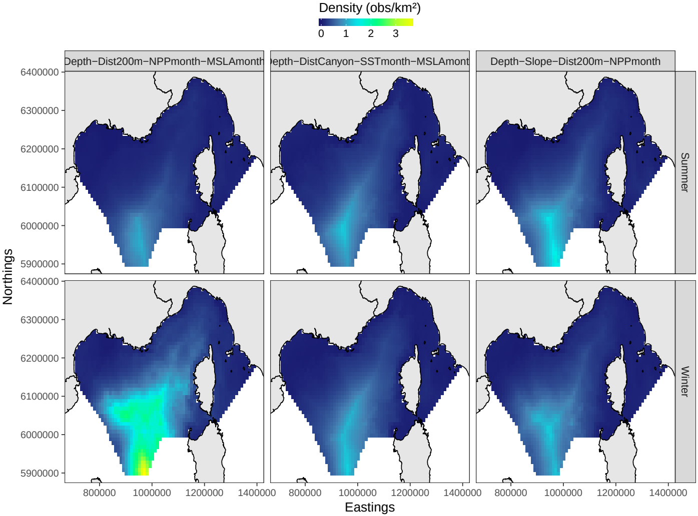


Out-of-Sample validation

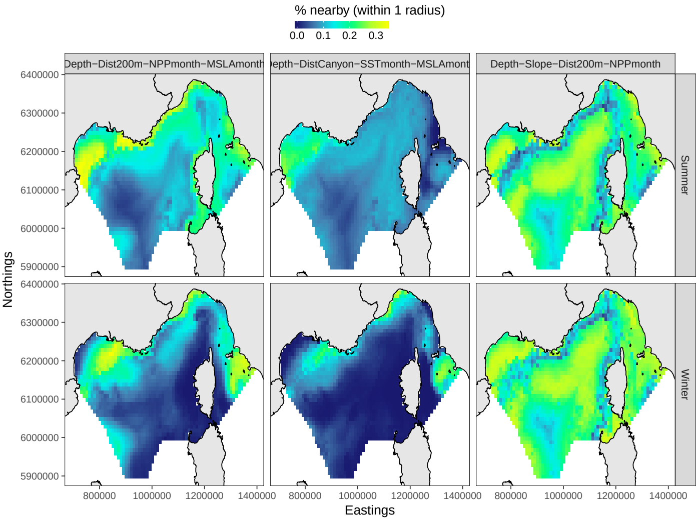
RMSE	${}_{0.387}0.493_{0.587}$	${}_{0.270}0.352_{0.402}$	${}_{0.314}0.360_{0.408}$
INT_{α}	271.5	174.5	159.0



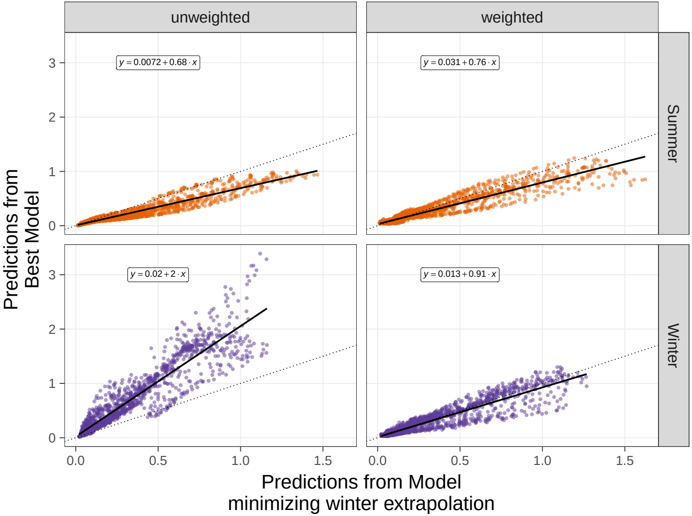
Predictions



Predictions



Predictions



Discussion

Statistical robustness

Robust statistics is an extension of parametric statistics, taking into account that parametric models are at best only approximations to reality. (Ronchetti, 2014)

Statistical robustness

Robust statistics is an extension of parametric statistics, taking into account that parametric models are at best only approximations to reality. (Ronchetti, 2014)

Robust statistical methods are procedures that give approximately the same results as classical methods when there are no atypical observations, and are only slightly affected by a small or moderate proportion of atypical observations. (Marrona, 2014)

Statistical robustness

Robust statistics is an extension of parametric statistics, taking into account that parametric models are at best only approximations to reality. (Ronchetti, 2014)

Robust statistical methods are procedures that give approximately the same results as classical methods when there are no atypical observations, and are only slightly affected by a small or moderate proportion of atypical observations. (Marrona, 2014)

Robustness primarily should be concerned with safeguarding against ill effects caused by finite but small deviations from an idealized model, with emphasis on the words small and model. (Huber, 2014)

Statistical robustness

Emphasis on (parametric) model (mis-)specification

Statistical robustness

Emphasis on (parametric) model (mis-)specification

what's "atypical", "small", "ill effects" . . . is not operationalized precisely

Statistical robustness

Emphasis on (parametric) model (mis-)specification

what's "atypical", "small", "ill effects" ... is not operationalized precisely

→ gives too much 'researcher degrees of freedom'? (Simmons et al., 2011)

Inferential brittleness? Predictive robustness?

Different paths to perform a specification search

$\rightarrow$ different inferences wrt to processes...

Inferential brittleness? Predictive robustness?

Different paths to perform a specification search

→ different inferences wrt to processes...

→ qualitative difference wrt to predictions (extra- *vs* inter-polations)...

Inferential brittleness? Predictive robustness?

Different paths to perform a specification search

$\longrightarrow$ different inferences wrt to processes...

$\longrightarrow$ qualitative difference wrt to predictions (extra- *vs* inter-polations)...

does not necessarily translate into quantitative differences!

Inferential brittleness? Predictive robustness?

Different paths to perform a specification search

→ different inferences wrt to processes...

→ qualitative difference wrt to predictions (extra- *vs* inter-polations)...

does not necessarily translate into quantitative differences!
Many-to-one mapping = Predictive Promiscuity

⇒ need for micro-foundations *sensu* Achen (2002)

Any role for this weighted likelihood approach?

Thanks & Questions, comments welcome

References

ACHEN, C. H. (2002). Toward a New Political Methodology: Microfoundations and ART. *Annual Review of Political Science* 5 423–450.

CARPENTER, B., GELMAN, A., HOFFMAN, M. D., LEE, D., GOODRICH, B., BETANCOURT, M., BRUBAKER, M., GUO, J., LI, P. & RIDDELL, A. (2017). `Stan`: A Probabilistic Programming Language. *Journal of Statistical Software* 76. URL `https://www.jstatsoft.org/article/view/v076i01`.

EILERS, P. & MARX, B. (2010). Splines, Knots, and Penalties. *WIREs Computational Statistics* 2 637–653.

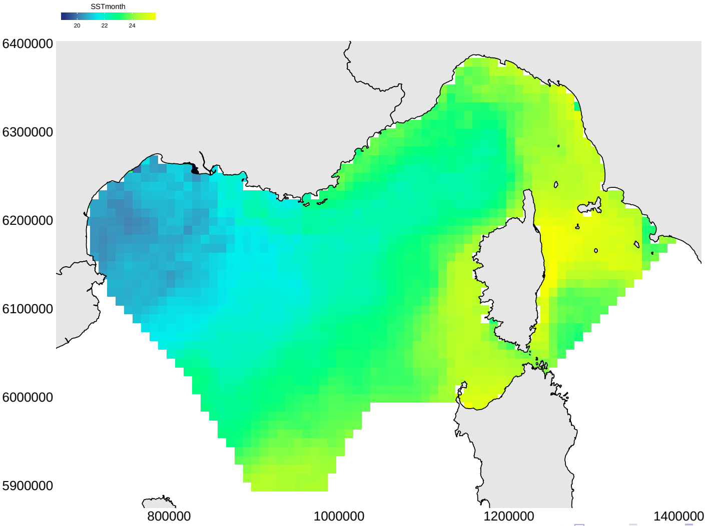
GRIFFIN, J. & BROWN, P. (2016). Hierarchical Shrinkage Priors for Regression Models. *Bayesian Analysis* 1–25.

HUBER, P. (2014). *International Encyclopedia of Statistical Science*, chap. Robust Statistics. Springer, 1248–1251.

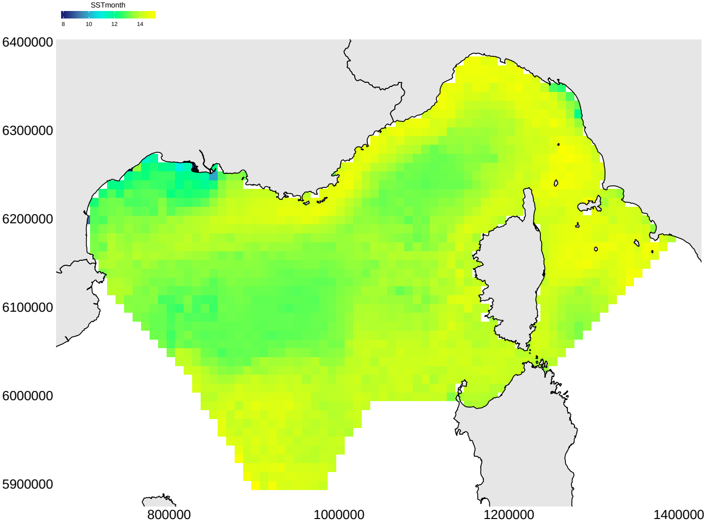
KING, G. & ZENG, L. (2007). When Can History Be Our Guide? The Pitfalls of Counterfactual Inference. *International Studies Quaterly* 51 183–210.

KLEIBER, C. & ZEILEIS, A. (2016). Visualizing Count Data Regression Using Rootograms. *The American Statistician* 70 296–303.

Extrapolation



Extrapolation



Extrapolation