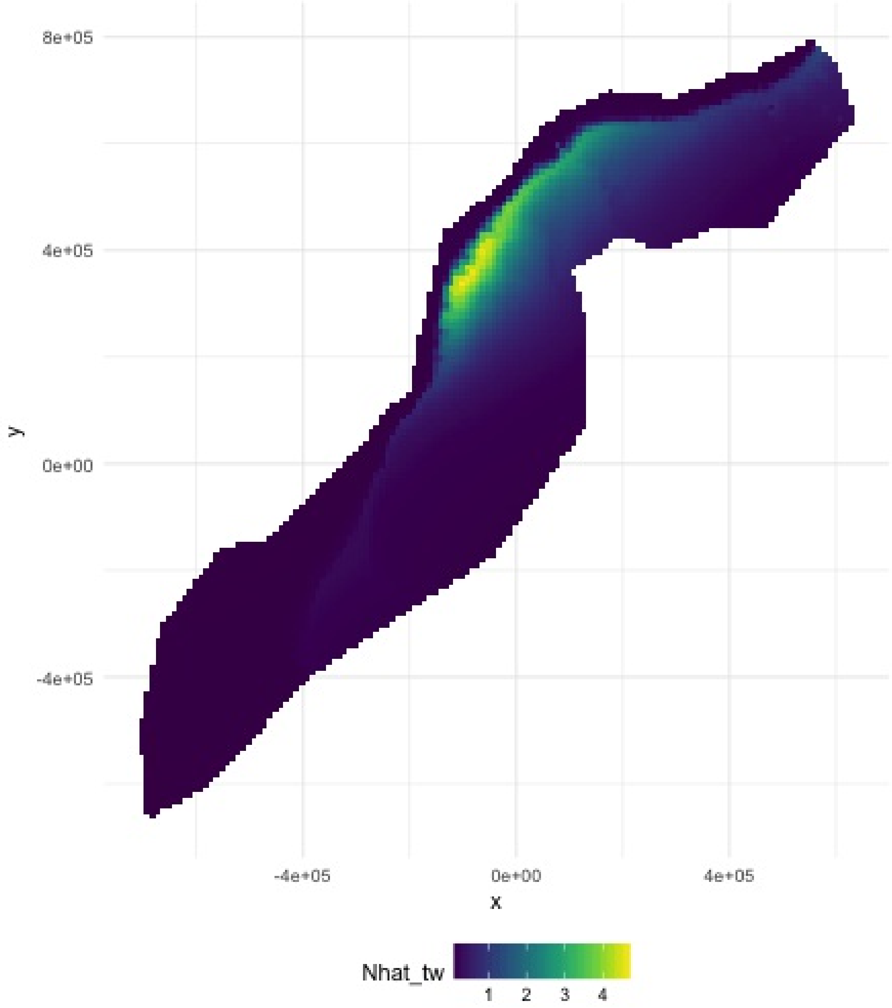
Making predictions

So far...

- Build, check & select models for detectability
- Build, check & select models for abundance
- Make some ecological inference about smooths
- **What about predictions?**

Let’s talk about maps

What does a map mean?



- Grids!
- Cells are abundance estimate
- “snapshot”
- Sum cells to get abundance
- Sum a subset?

Going back to the formula

(Count) Model:

$$n_j = A_j \hat{p}_j \exp\left[\beta_0 + s(\mathrm{y}_j) + s(\mathrm{Depth}_j)\right] + \epsilon_j$$

Predictions (index r):

$$n_r = A_r \exp[\beta_0 + s(\mathrm{y}_r) + s(\mathrm{Depth}_r)]$$

Need to “fill-in” values for A_r, y_r and Depth_r.

Predicting

- With these values can use predict in R
- predict(model, newdata=data)

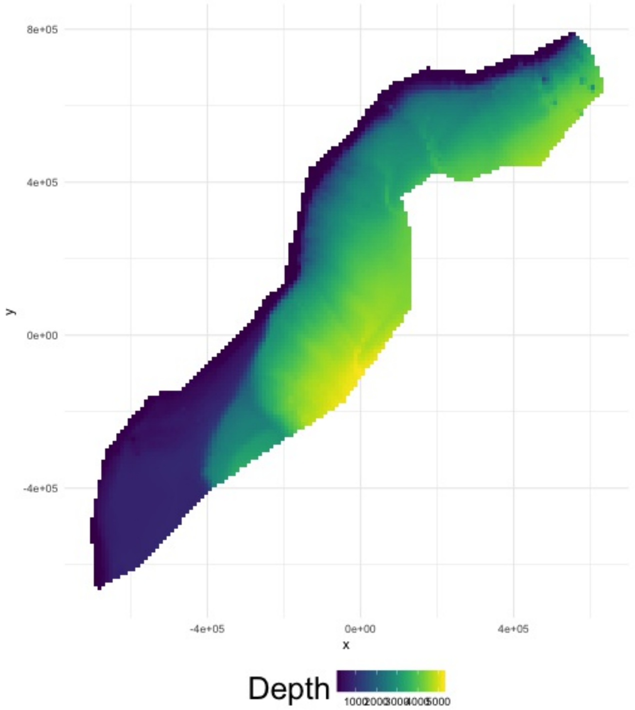
Prediction data

```
           x      y     Depth    SST      NPP DistToCAS     EKE off.set
126 547984.6 788254  153.5983 8.8812 1462.521 11788.974 0.0074   1e+08
127 557984.6 788254  552.3107 9.2078 1465.410  5697.248 0.0144   1e+08
258 527984.6 778254   96.8199 9.6341 1429.432 13722.626 0.0024   1e+08
259 537984.6 778254  138.2376 9.6650 1424.862  9720.671 0.0027   1e+08
260 547984.6 778254  505.1439 9.7905 1379.351  8018.690 0.0101   1e+08
261 557984.6 778254 1317.5952 9.9523 1348.544  3775.462 0.0193   1e+08
    LinkID    Nhat_tw
126      1 0.01417657
127      2 0.05123483
258      3 0.01118858
259      4 0.01277096
260      5 0.04180434
261      6 0.45935801
```

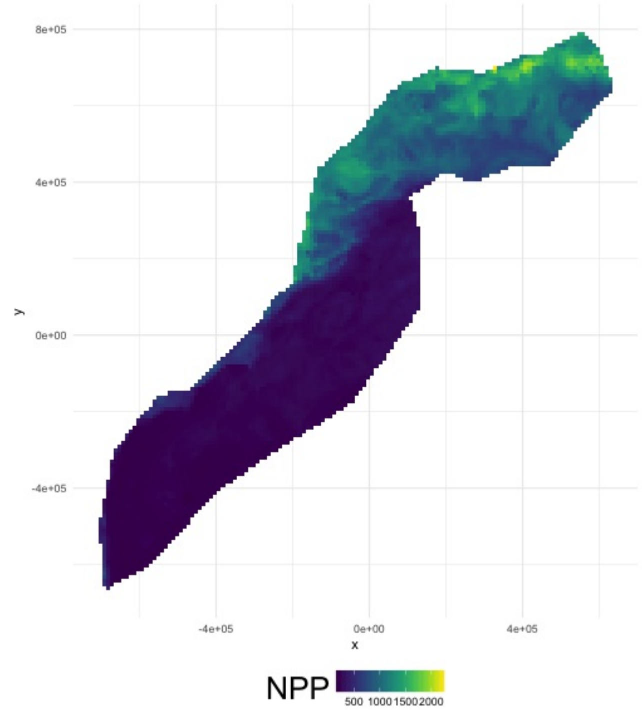
A quick word about rasters

- We have talked about rasters a bit
- In R, the data.frame is king
- Fortunately as.data.frame exists
- Make our "stack" and then convert to data.frame

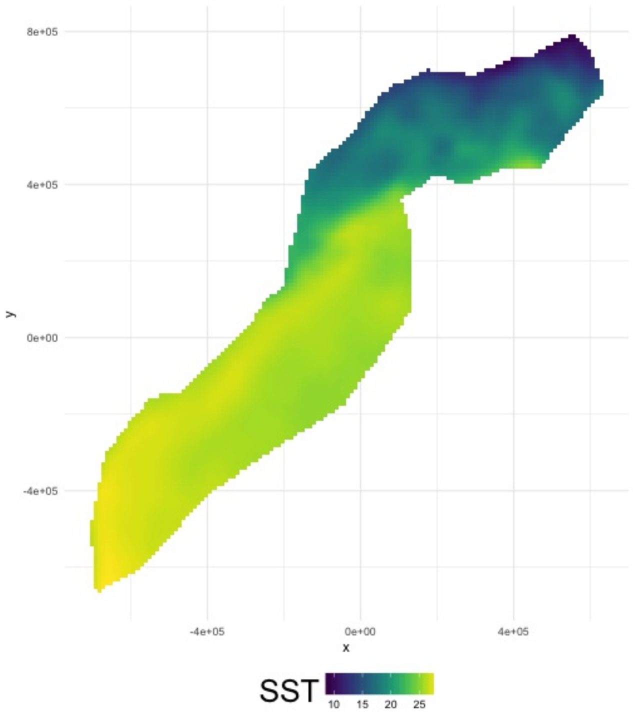
Predictors



8e+05
4e+05
0e+00
-4e+05
y
-4e+05
0e+00
4e+05
x
Depth
1000 2000 3000 4000 5000



8e+05
4e+05
0e+00
-4e+05
y
-4e+05
0e+00
4e+05
x
NPP
500 1000 1500 2000



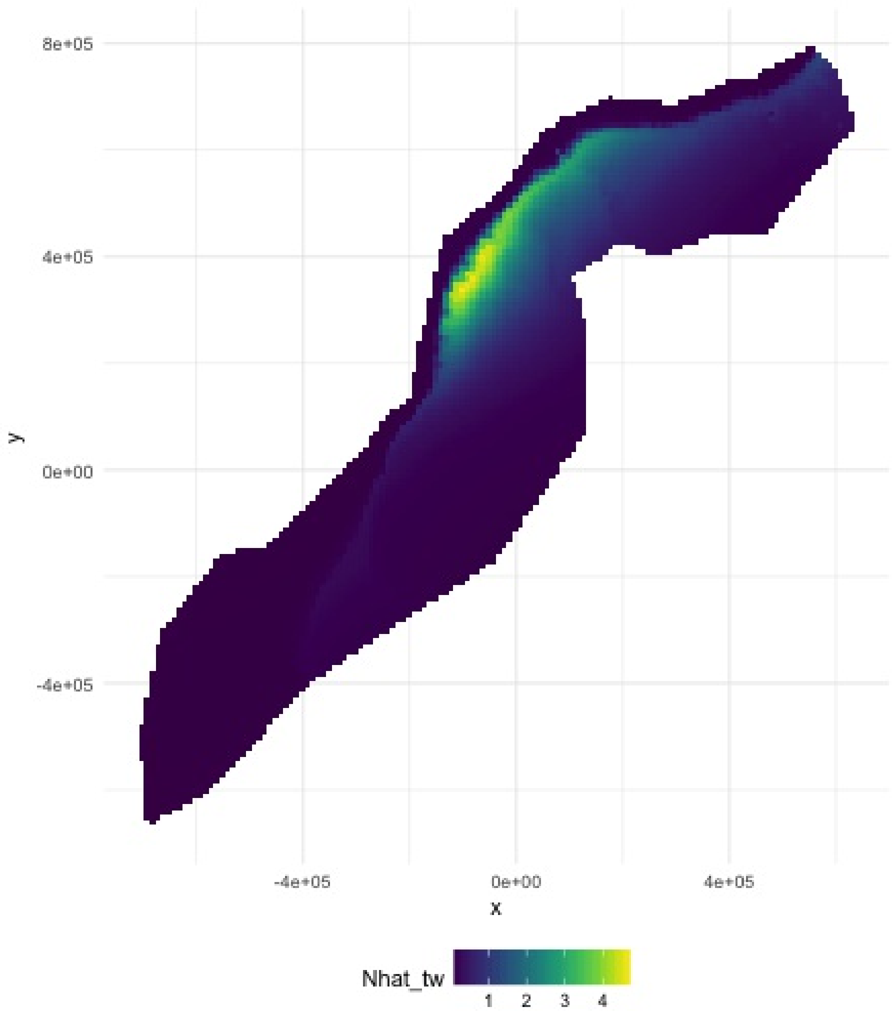
8e+05
4e+05
0e+00
-4e+05
y
-4e+05
0e+00
4e+05
x
SST
10 15 20 25

Making a prediction

- Add another column to the prediction data
- Plotting then easier (in R)

```
predgrid$Nhat_tw <- predict(dsm_all_tw_rm, predgrid)
```

Maps of predictions




```
p <- ggplot(predgrid) +
      geom_tile(aes(x=x, y=y,
                    fill=Nhat_tw)) +
      scale_fill_viridis() +
      coord_equal()
print(p)
```

Total abundance

Each cell has an abundance, sum to get total

```
sum(predict(dsm_all_tw_rm, predgrid))
```

```
[1] 2491.864
```

Subsetting

R subsetting lets you calculate “interesting” estimates:

```
# how many sperm whales at depths less than 2500m?
sum(predgrid$Nhat_tw[predgrid$Depth < 2500])
```

```
[1] 1006.272
```

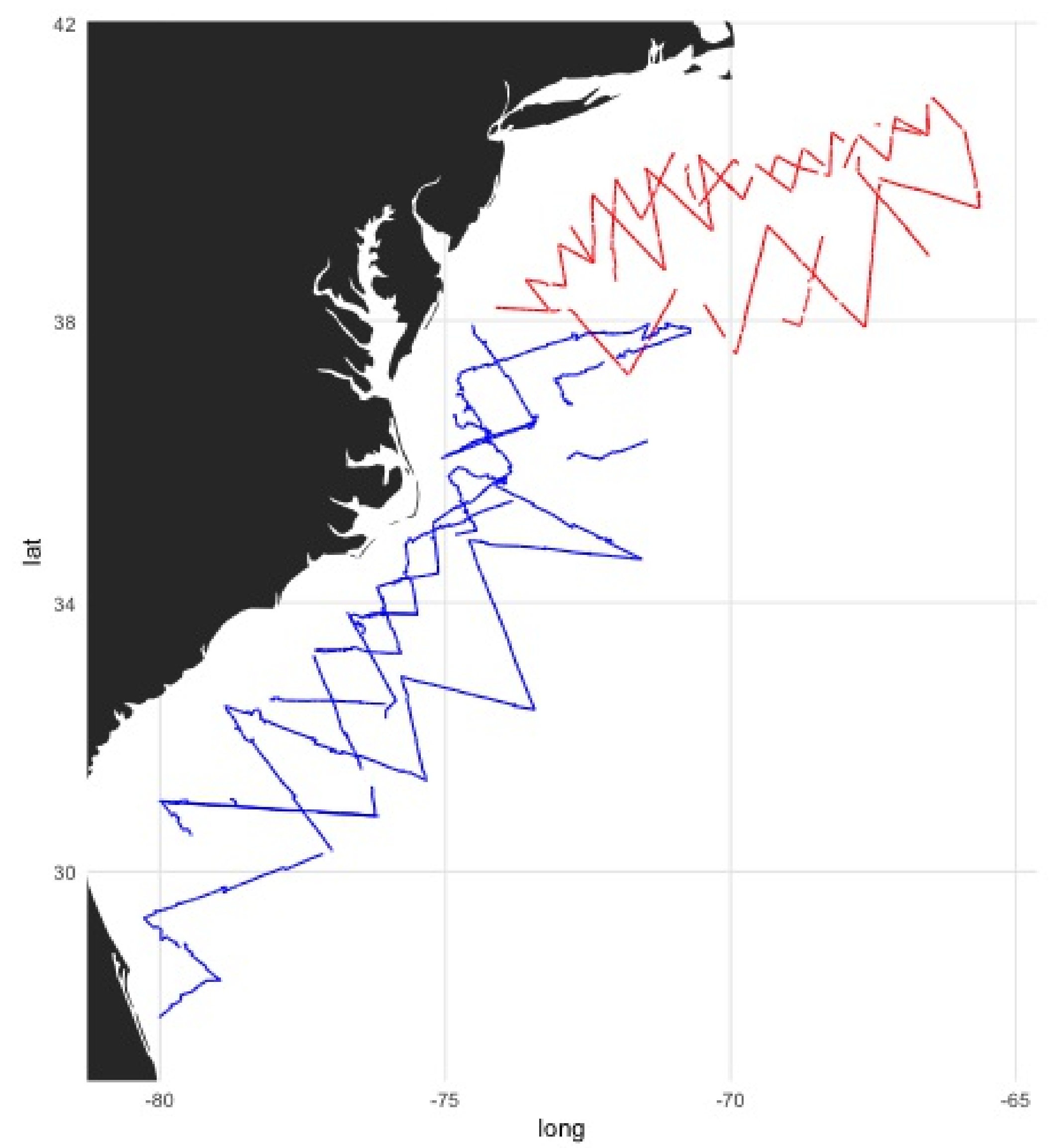
```
# how many sperm whales North of 0?
sum(predgrid$Nhat_tw[predgrid$x>0])
```

```
[1] 1383.742
```

Extrapolation

What do we mean by extrapolation?

- Predicting at values outside those observed
- What does "outside" mean?
  - between transects?
  - outside "survey area"?

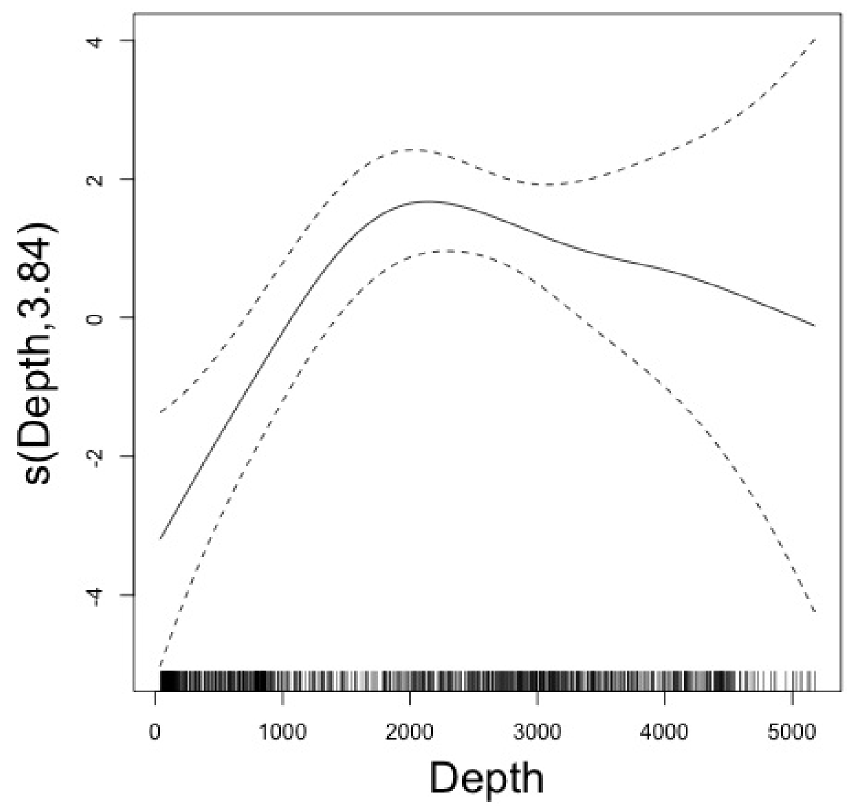


Temporal extrapolation

- Models are temporally implicit (mostly)
- Dynamic variables change seasonally
- Migration can be an issue
- Need to understand what the predictions **are**

Extrapolation

- Extrapolation is fraught with issues
- Want to be predicting “inside the rug”
- In general, try not to do it!
- (Think about variance too!)



Recap

- Using predict
- Getting “overall” abundance
- Subsetting
- Plotting in R
- Extrapolation (and its dangers)

Estimating variance

Now we can make predictions

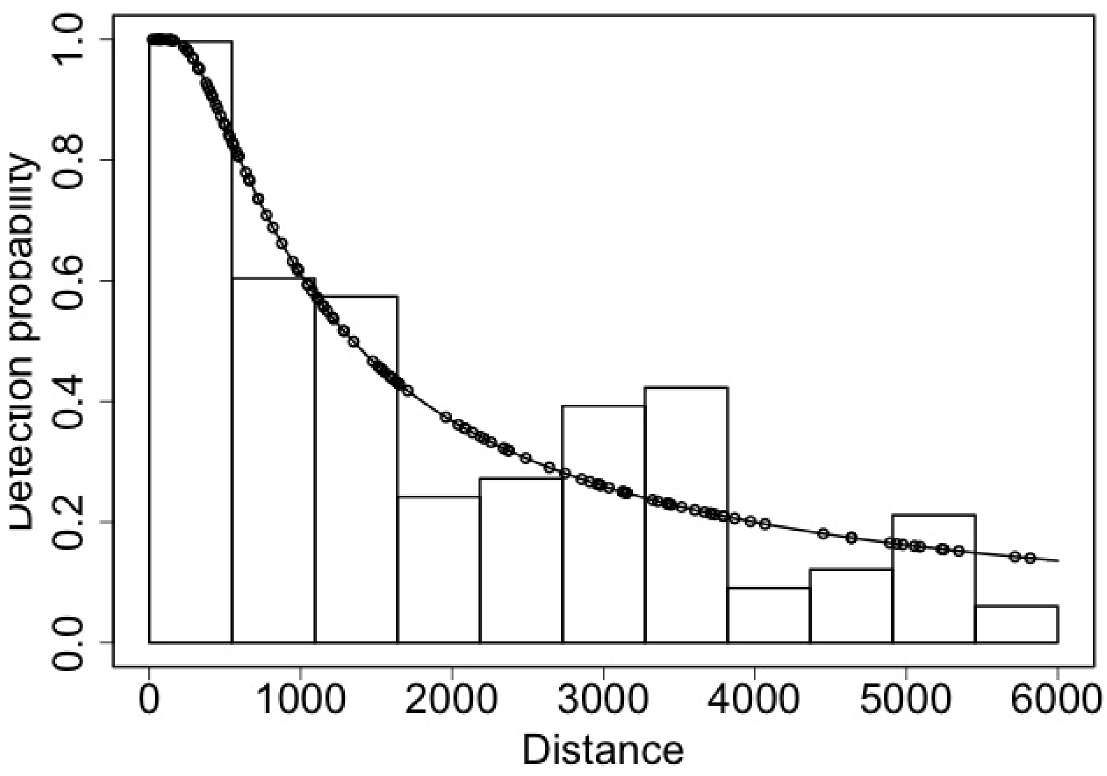
Now we are dangerous.

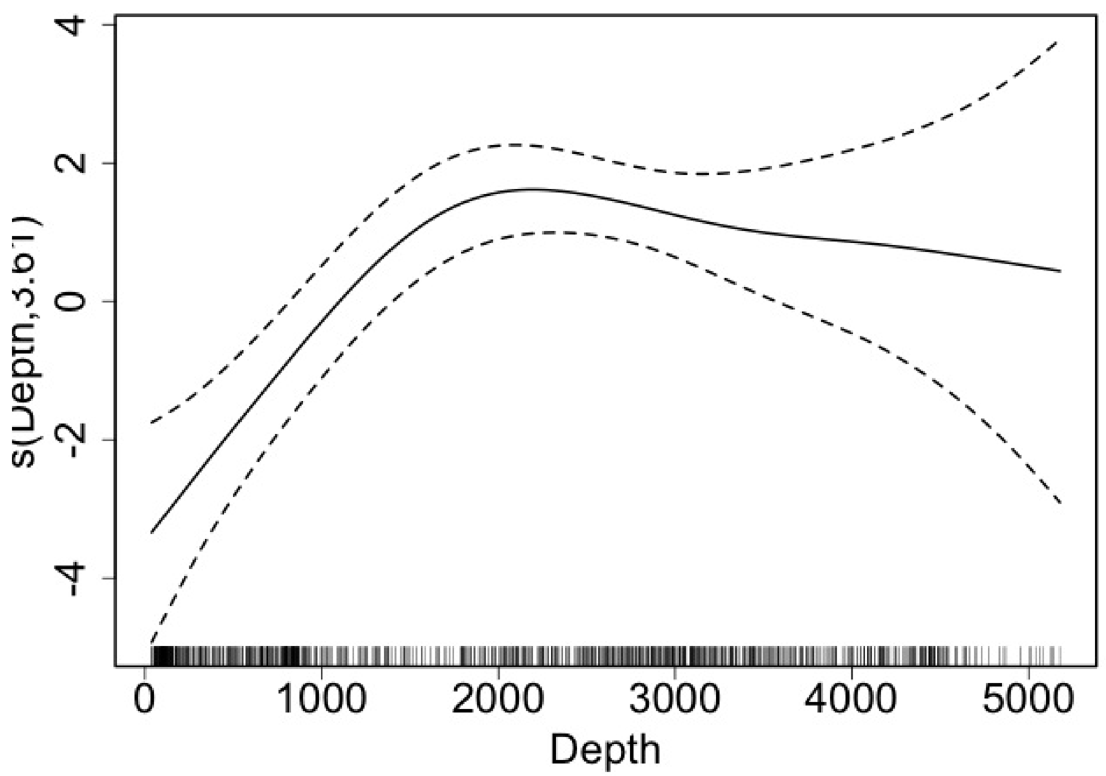
Predictions are useless without uncertainty

Where does uncertainty come from?

Sources of uncertainty

- Detection function
- GAM parameters

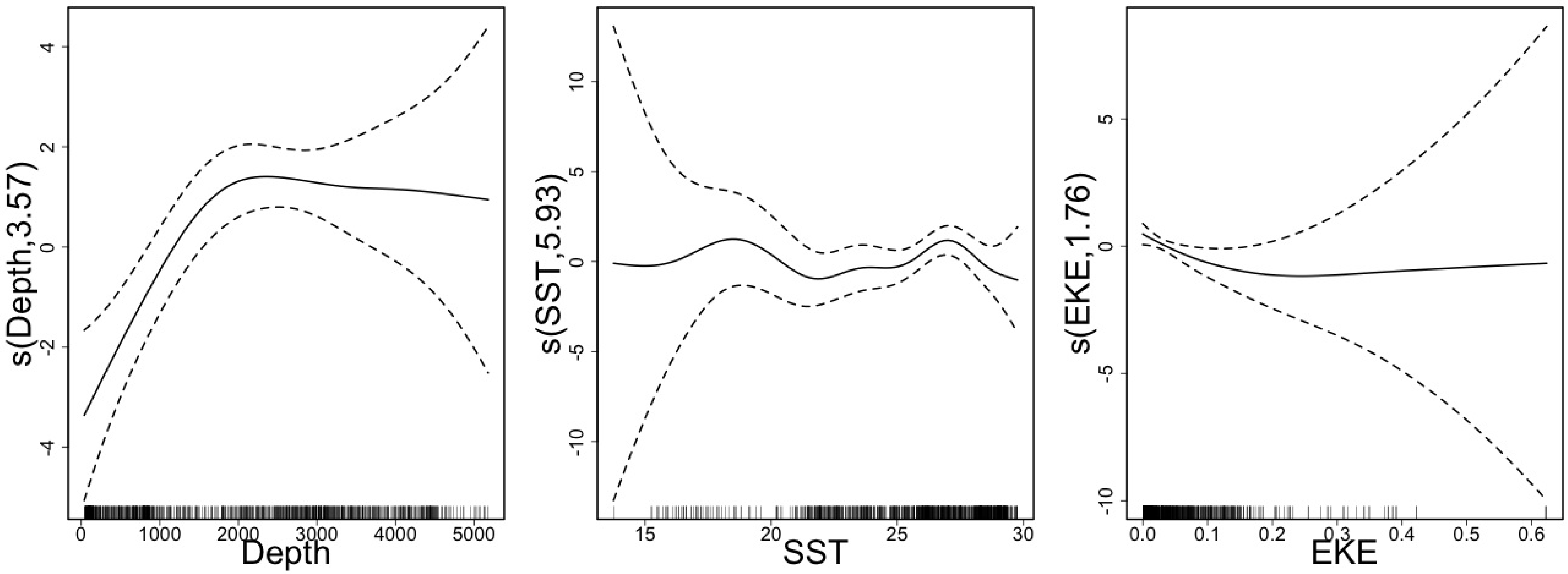




Let’s think about smooths first

Uncertainty in smooths

- Dashed lines are +/- 2 standard errors
- How do we translate to $\hat{N}$?



Back to bases

- Before we expressed smooths as:
  - $s(x) = \sum_{k=1}^{K} \beta_k b_k(x)$
- Theory tells us that:
  - $\boldsymbol{\beta} \sim N(\hat{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}})$
  - where $\mathbf{V}_{\boldsymbol{\beta}}$ is a bit complicated
  - (derived from the smoother matrix)

Predictions to prediction variance (roughly)

- “map” data onto fitted values $\mathbf{X}\boldsymbol{\beta}$
- “map” prediction matrix to predictions $\mathbf{X}_p\boldsymbol{\beta}$
- Here $\mathbf{X}_p$ need to take smooths into account
- pre-/post-multiply by $\mathbf{X}_p$ to “transform variance”
  - $\Rightarrow \mathbf{X}_p^{\mathrm{T}}\mathbf{V}_{\boldsymbol{\beta}}\mathbf{X}_p$
  - link scale, need to do another transform for response

Adding in detection functions

GAM + detection function uncertainty

(Getting a little fast-and-loose with the mathematics)

$$\text{CV}^2\left(\hat{N}\right) \approx \text{CV}^2\left(\text{GAM}\right) + \text{CV}^2\left(\text{detection function}\right)$$

Not that simple...

- Assumes detection function and GAM are **independent**
- **Maybe** this is okay?
- (Probably not true?)

Variance propagation

- Include the detectability as term in GAM
- Random effect, mean zero, variance of detection function
- Uncertainty "propagated" through the model
- Details in bibliography (too much to detail here)
- Under development
- (Can cover in special topic)

That seemed complicated…

R to the rescue

In R...

- Functions in dsm to do this
- `dsm.var.gam`
  - assumes spatial model and detection function are independent
- `dsm.var.prop`
  - propagates uncertainty from detection function to spatial model
  - only works for count models (more or less)

Variance of abundance

Using dsm.var.gam

```r
dsm_tw_var_ind <- dsm.var.gam(dsm_all_tw_rm, predgrid,
                              off.set=predgrid$off.set)
summary(dsm_tw_var_ind)
```

```
Summary of uncertainty in a density surface model calculated
 analytically for GAM, with delta method

Approximate asymptotic confidence interval:
    2.5%     Mean    97.5%
1539.018 2491.864 4034.643
(Using log-Normal approximation)

Point estimate                 : 2491.864
CV of detection function       : 0.2113123
CV from GAM                    : 0.1329
Total standard error           : 622.0389
Total coefficient of variation : 0.2496
```

Variance of abundance

Using dsm.var.prop

```
dsm_tw_var <- dsm.var.prop(dsm_all_tw_rm, predgrid,
                           off.set=predgrid$off.set)
summary(dsm_tw_var)
```

```
Summary of uncertainty in a density surface model calculated
 by variance propagation.

Probability of detection in fitted model and variance model
  Fitted.model Fitted.model.se Refitted.model
1    0.3624567      0.07659373      0.3624567

Approximate asymptotic confidence interval:
    2.5%     Mean    97.5%
1556.898 2458.634 3882.646
(Using log-Normal approximation)

Point estimate                 : 2458.634
Standard error                 : 581.0379
Coefficient of variation       : 0.2363
```

Plotting - data processing

- Calculate uncertainty per-cell
- dsm.var.* thinks predgrid is one “region”
- Need to split data into cells (using split())
- (Could be arbitrary sets of cells, see exercises)
- Need width and height of cells for plotting

Plotting (code)

```
predgrid$width <- predgrid$height <- 10*1000
predgrid_split <- split(predgrid, 1:nrow(predgrid))
head(predgrid_split,3)
```

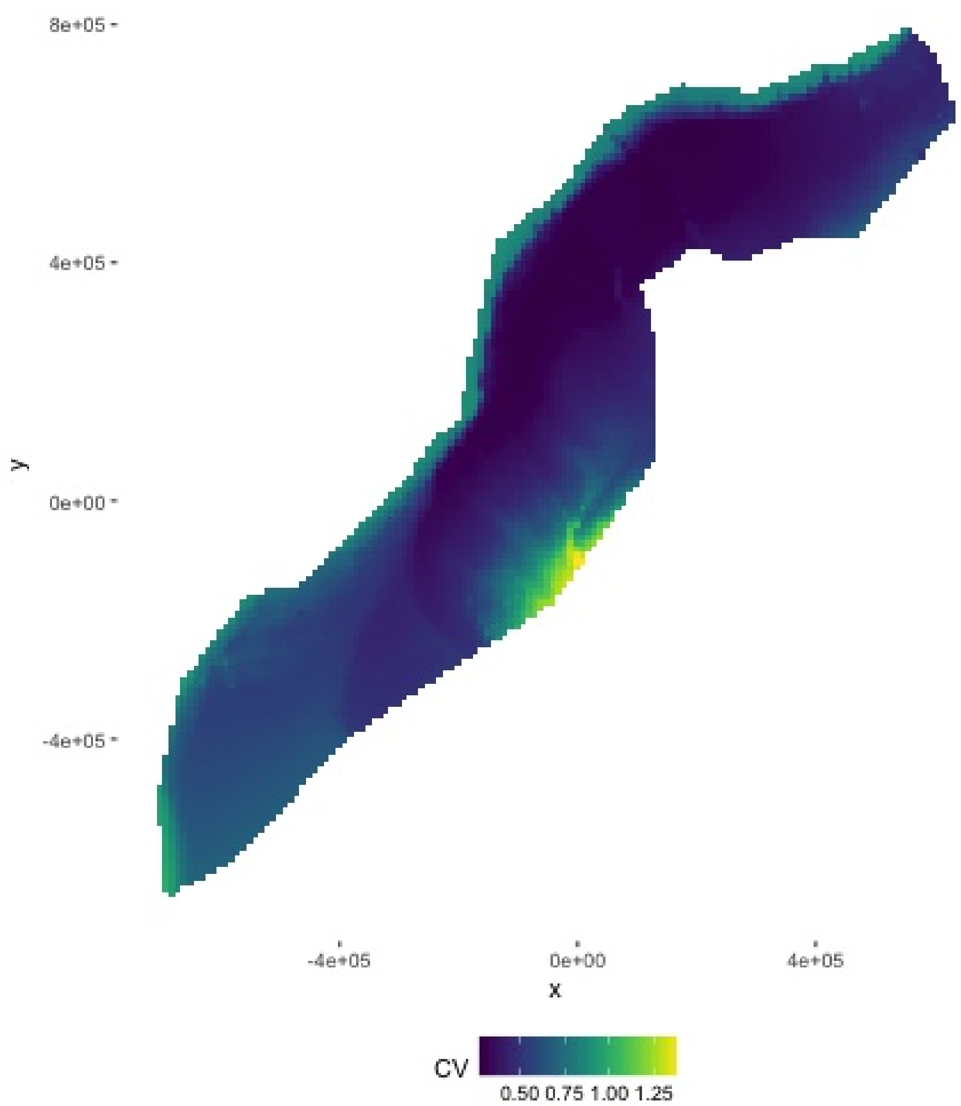
```
$`1`
           x      y    Depth    SST      NPP DistToCAS    EKE off.set
126 547984.6 788254 153.5983 8.8812 1462.521  11788.97 0.0074   1e+08
    LinkID    Nhat_tw height width
126      1 0.01417657  10000 10000

$`2`
           x      y    Depth    SST     NPP DistToCAS    EKE off.set
127 557984.6 788254 552.3107 9.2078 1465.41  5697.248 0.0144   1e+08
    LinkID    Nhat_tw height width
127      2 0.05123483  10000 10000

$`3`
           x      y   Depth    SST     NPP DistToCAS    EKE off.set
258 527984.6 778254 96.8199 9.6341 1429.432  13722.63 0.0024   1e+08
    LinkID    Nhat_tw height width
258      3 0.01118858  10000 10000
```

```
dsm_tw_var_map <- dsm.var.prop(dsm_all_tw_rm, predgrid_split,
                               off.set=predgrid$off.set)
```

CV plot

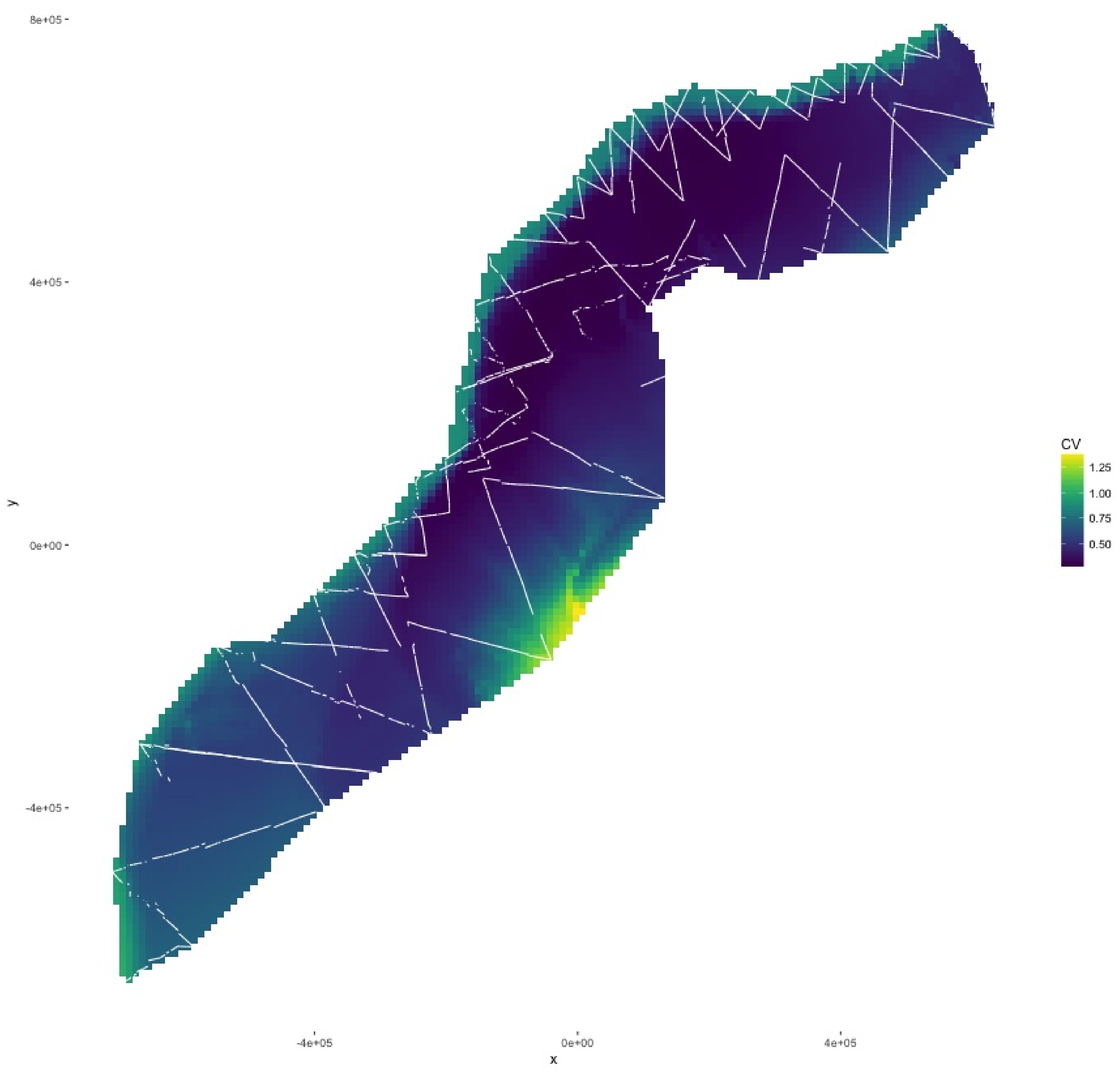



```
p <- plot(dsm_tw_var_map,
observations=FALSE, plot=FALSE) +
      coord_equal() +
      scale_fill_viridis()
print(p)
```

Interpreting CV plots

- Plotting coefficient of variation
- Standardise standard deviation by mean
- $\text{CV} = \text{se}(\hat{N})/\hat{N}$ (per cell)
- Can be useful to overplot survey effort

Effort overplotted



Big CVs

- Here CVs are "well behaved"
- Not always the case (huge CVs possible)
- These can be a pain to plot
- Use cut() in R to make categorical variable
  - e.g. c(seq(0,1, len=100), 2:4, Inf) or somesuch

Recap

- How does uncertainty arise in a DSM?
- Estimate variance of abundance estimate
- Map coefficient of variation

Let’s try that!