Lecture 5: Predictions and variance

University of St Andrews

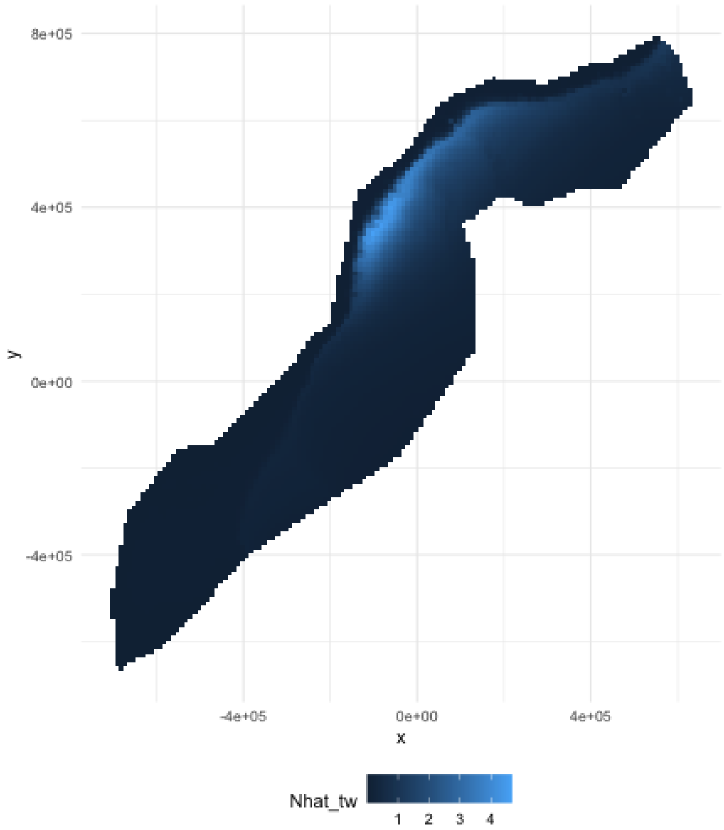
So far...

- Build, check & select *detection* models
- Build, check & select *spatial* models

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 (j observations):

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

Predictions (index r):

$$\hat{n}_r = A_r \exp\left[\hat{\beta}_0 + \hat{s}(\text{y}_r) + \hat{s}(\text{Depth}_r)\right]$$

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

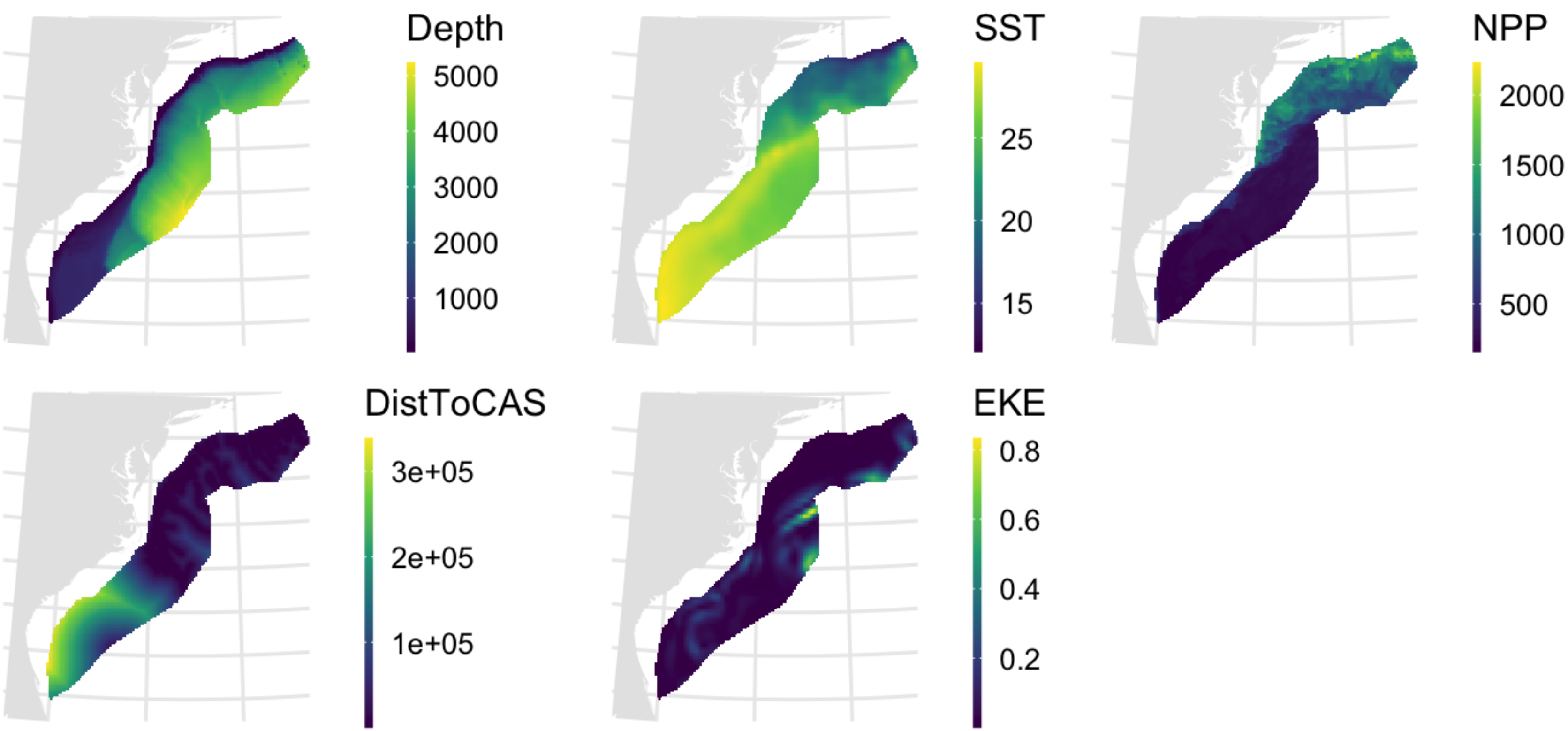
Predicting

- With these values can use `predict` in R
- `predict(model, newdata=data, off.set=off.set)`
- `off.set` gives the area of the grid cells
- more info in `?predict.dsm`

Prediction data

```
##              x      y      Depth      SST      NPP DistToCAS
## 126 547984.6 788254  153.59825 12.04609 1462.521 11788.974
## 127 557984.6 788254  552.31067 12.81379 1465.410  5697.248
## 258 527984.6 778254   96.81992 12.90251 1429.432 13722.626
## 259 537984.6 778254  138.23763 13.21393 1424.862  9720.671
## 260 547984.6 778254  505.14386 13.75655 1379.351  8018.690
## 261 557984.6 778254 1317.59521 14.42525 1348.544  3775.462
##                EKE off.set      long      lat
## 126 0.0008329031   1e+08 -66.52252 40.94697
## 127 0.0009806611   1e+08 -66.40464 40.94121
## 258 0.0011575423   1e+08 -66.76551 40.86781
## 259 0.0013417297   1e+08 -66.64772 40.86227
## 260 0.0026881567   1e+08 -66.52996 40.85662
## 261 0.0045683752   1e+08 -66.41221 40.85087
```

Predictors

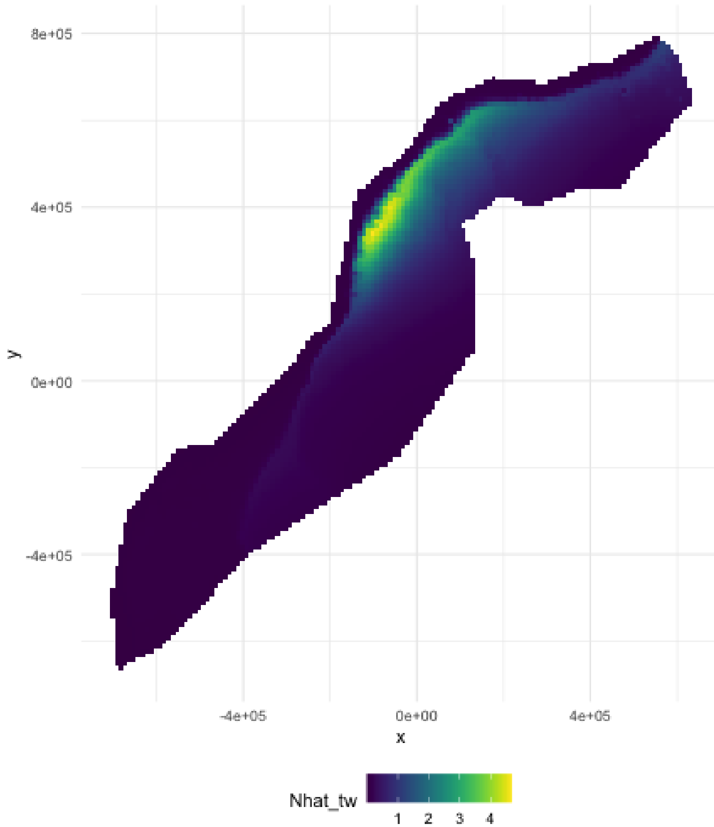

Depth
5000
4000
3000
2000
1000
SST
25
20
15
NPP
2000
1500
1000
500
DistToCAS
3e+05
2e+05
1e+05
EKE
0.8
0.6
0.4
0.2

Making a prediction

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

```
predgrid$Nhat_tw <- predict(dsm_all_tw_rm,
                            predgrid,
                            off.set=predgrid$off.set)
```

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(predgrid$Nhat_tw)
```

```
## [1] 2491.863
```

Subsetting

R subsetting lets you calculate "interesting" estimates:

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

```
## [1] 1006.27
```

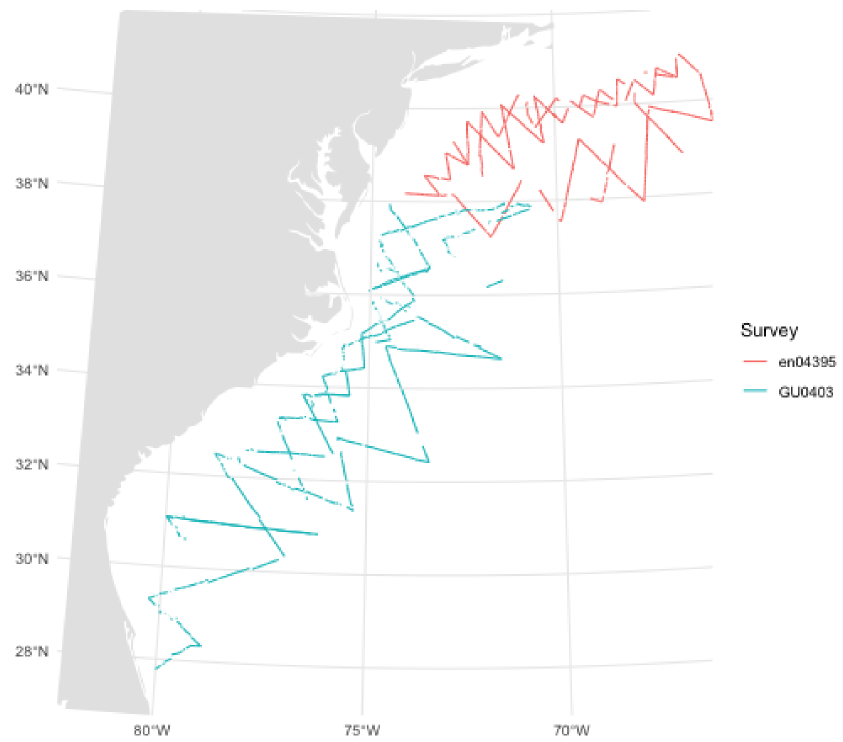
```
# how many sperm whales East of 0?
sum(predgrid$Nhat_tw[predgrid$x>0])
```

```
## [1] 1383.744
```

Extrapolation

What do we mean by extrapolation?

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



Extrapolation

- In general, try not to do it!
- Variance issues?
- Space-time interchangability?
- `dsmextra` package by Phil Bouchet
  - https://densitymodelling.github.io/dsmextra/index.html

Prediction recap

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

Estimating variance

Now we can make predictions

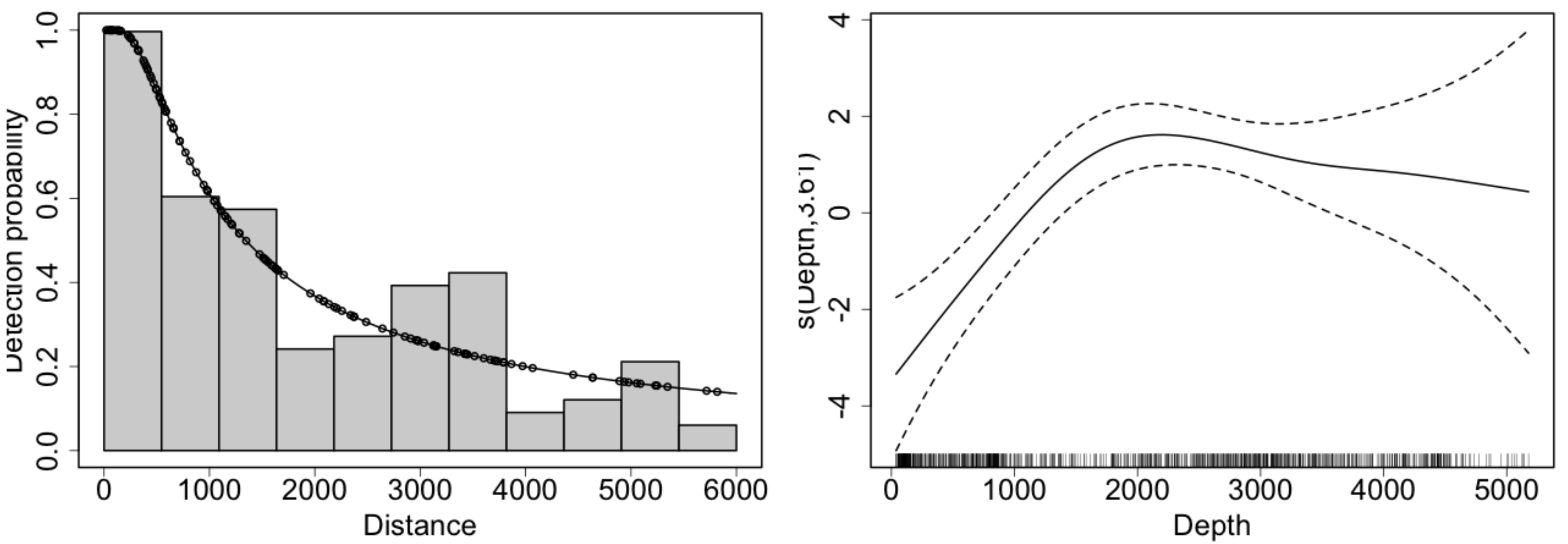
Now we are dangerous.

Predictions are useless without uncertainty

Where does uncertainty come from?

Sources of uncertainty

- Detection function parameters
- GAM parameters
- (And more! But only looking at these 2 here!)



Uncertianty of what?

- Uncertainty from detection function + GAM
- Want to talk about $\hat{N}$, so need to do some maths
- `dsm` does this for you!
- Details in Miller et al (2013) appendix

GAM + detection function uncertainty

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

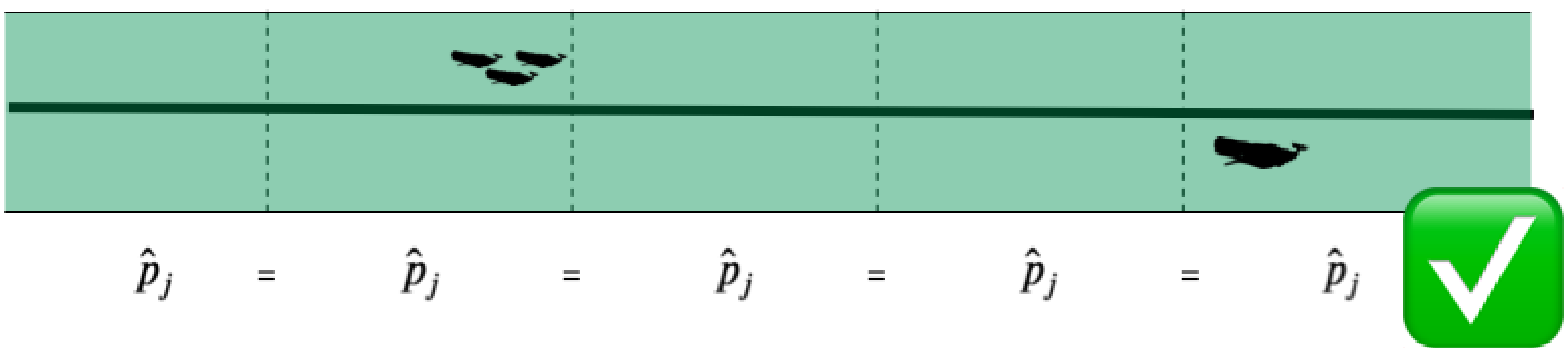
$$\mathrm{CV}^2\left(\hat{N}\right) \approx \mathrm{CV}^2\left(\mathrm{GAM}\right) +$$

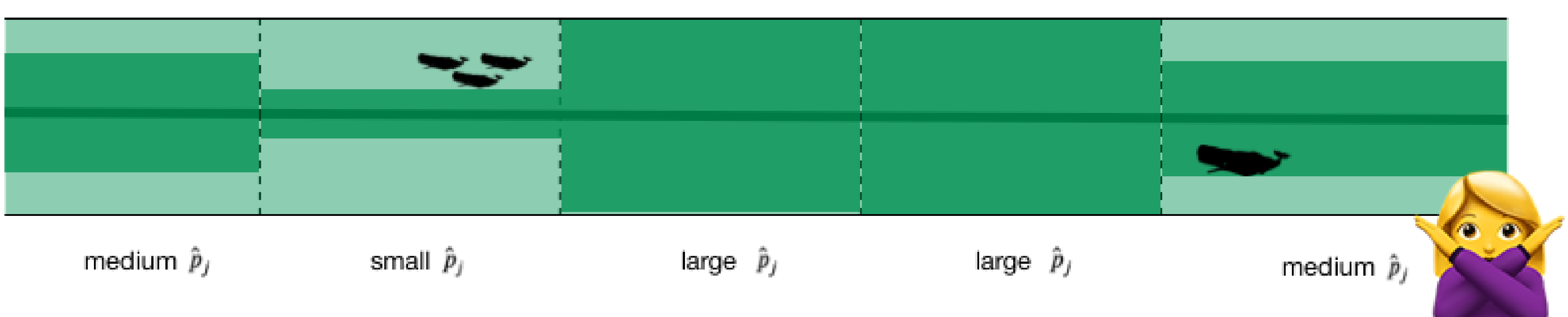
$$\mathrm{CV}^2\left(\text{detection function}\right)$$

the "delta method"

When can we use the delta method?

- Assumes detection function and GAM are **independent**
- This is okay if:
  - no detection function covariates





Variance propagation

- When detection function is not independent
- Uncertainty "propagated" through the model
- Refit both models together
- Bravington, Miller and Hedley (2019)
  - https://arxiv.org/abs/1807.07996

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
    - covariates can only vary at segment level

Variance of abundance

Using `dsm.var.gam`

```
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%
## 1539.017 2491.863 4034.641
## (Using log-Normal approximation)
##
## Point estimate                 : 2491.863
## CV of detection function       : 0.2113123
## CV from GAM                    : 0.1329
## Total standard error           : 622.0386
## Total coefficient of variation : 0.2496
```

Plotting - data processing

- Calculate uncertainty per-cell
- `dsm.var.*` thinks `predgrid` is one "region"
- Need to split data into cells (using `split()`)
- Need `width` and `height` of cells for plotting

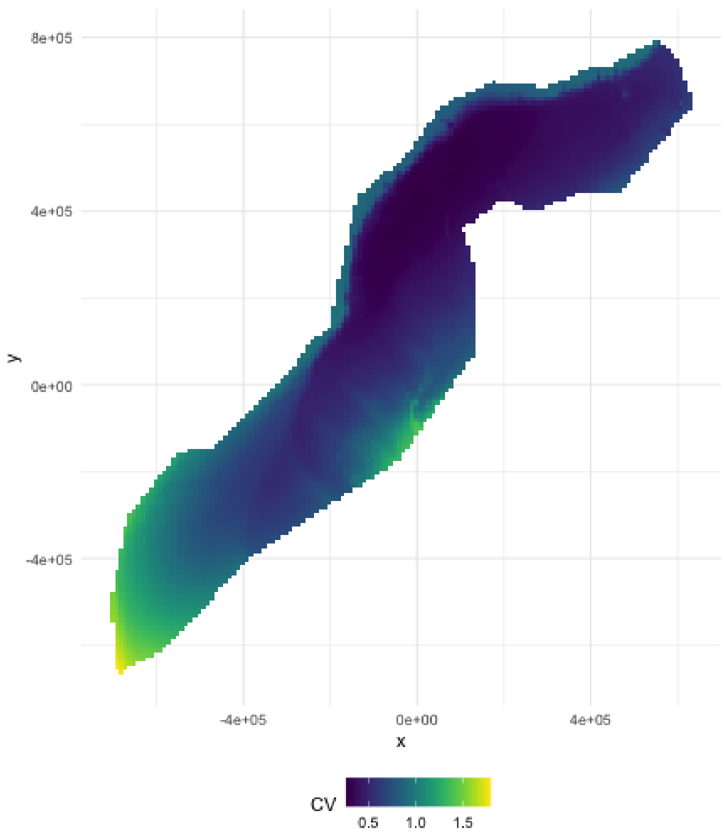
Plotting (code)

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

```
## $`1`
##            x      y    Depth      SST      NPP DistToCAS
## 126 547984.6 788254 153.5983 12.04609 1462.521  11788.97
##              EKE off.set      long      lat    Nhat_tw
## 126 0.0008329031   1e+08 -66.52252 40.94697 0.01417646
##     height width
## 126  10000 10000
##
## $`2`
##            x      y    Depth      SST     NPP DistToCAS
## 127 557984.6 788254 552.3107 12.81379 1465.41  5697.248
##              EKE off.set      long      lat    Nhat_tw
## 127 0.0009806611   1e+08 -66.40464 40.94121 0.05123446
##     height width
## 127  10000 10000
##
## $`3`
##            x      y    Depth      SST      NPP DistToCAS
## 258 527984.6 778254 96.81992 12.90251 1429.432  13722.63
##             EKE off.set      long      lat    Nhat_tw
```

CV plot

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

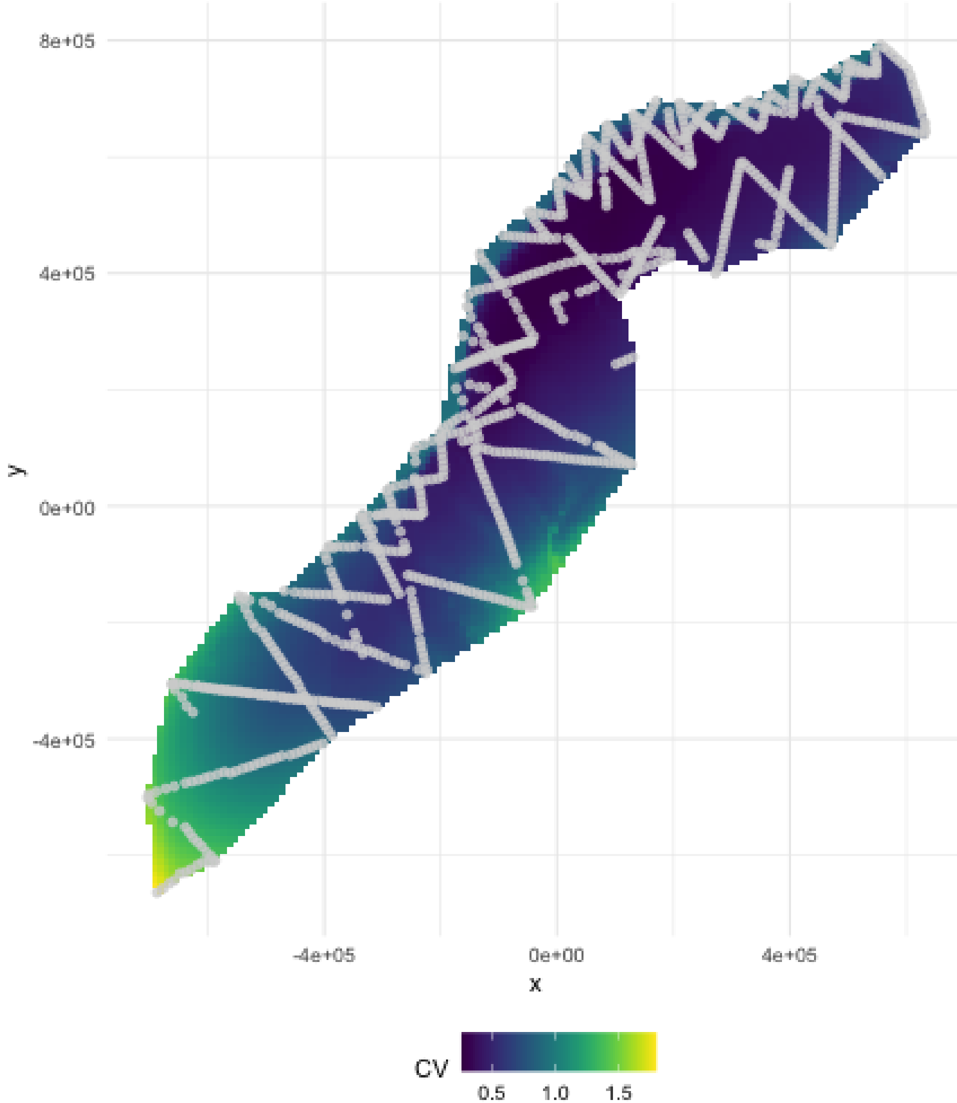



```
p <- plot(dsm_tw_var_map,
          observations=FALSE,
          plot=FALSE) +
      coord_equal() +
      scale_fill_viridis()
print(p)
```

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=10), 2:4, Inf)` or somesuch
- (Example in practical)

Uncertainty recap

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

Practical advice

Pilot studies and "you get what you pay for"

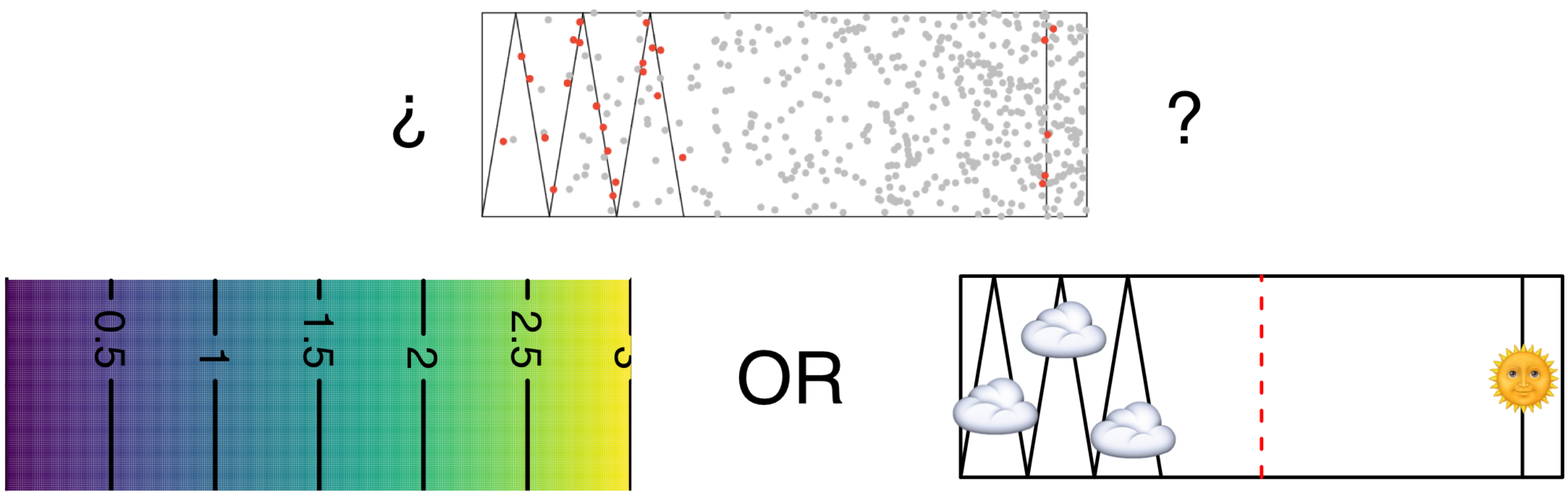
- Designing surveys is hard
- Designing surveys is essential
- Better to fail one season than fail for 5, 10 years
- Get information early, get it cheap
  - Inform design from a pilot study

Avoiding rules of thumb

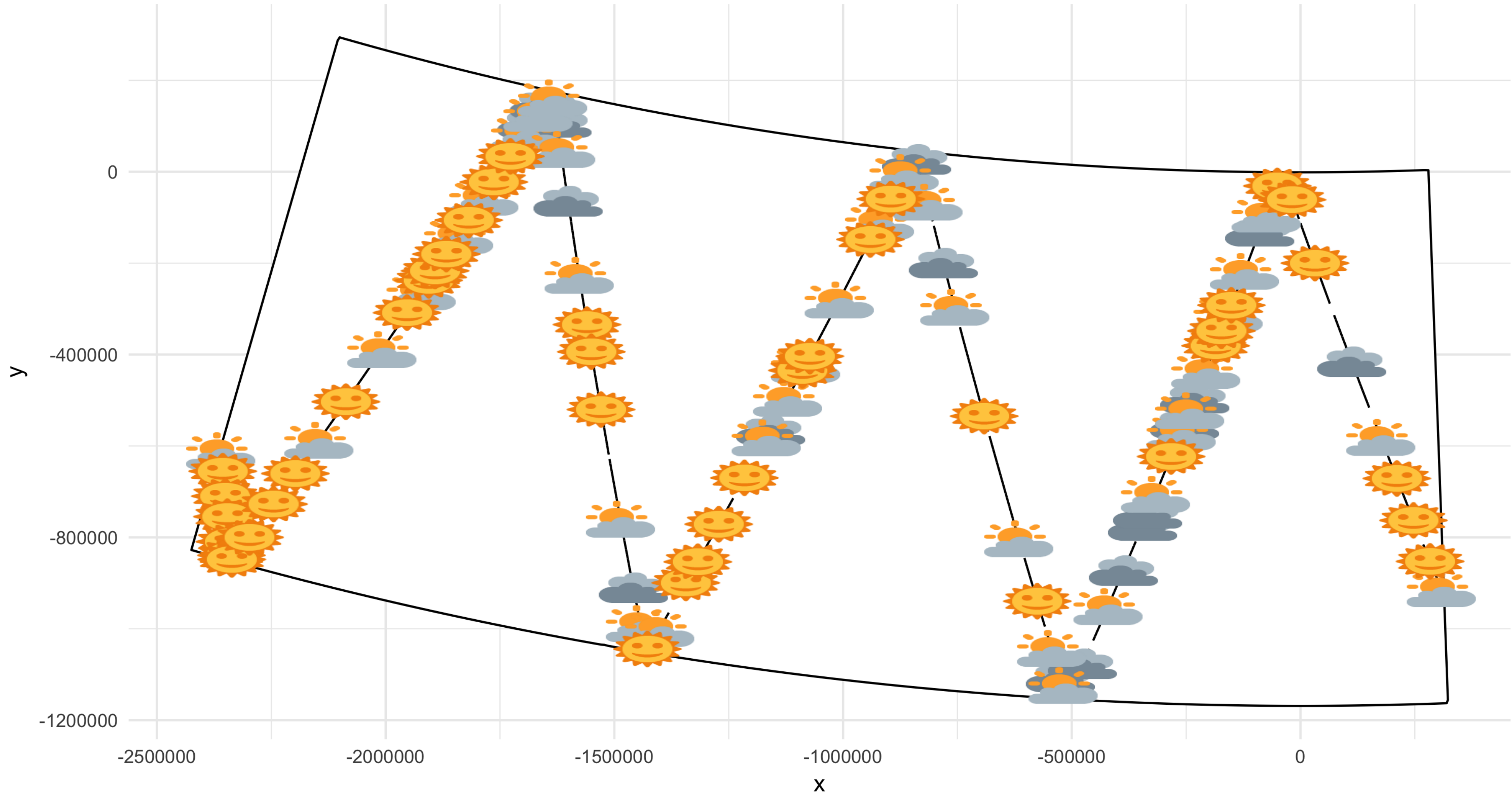
- Think about assumptions
  - Detection function
  - Spatial model
- Think about design
  - Spatial coverage
  - Covariate coverage

Sometimes things are complicated

- Weather has a big effect on detectability
- Need to record during survey
- Disambiguate between distribution/detectability
- Potential confounding can be BAD



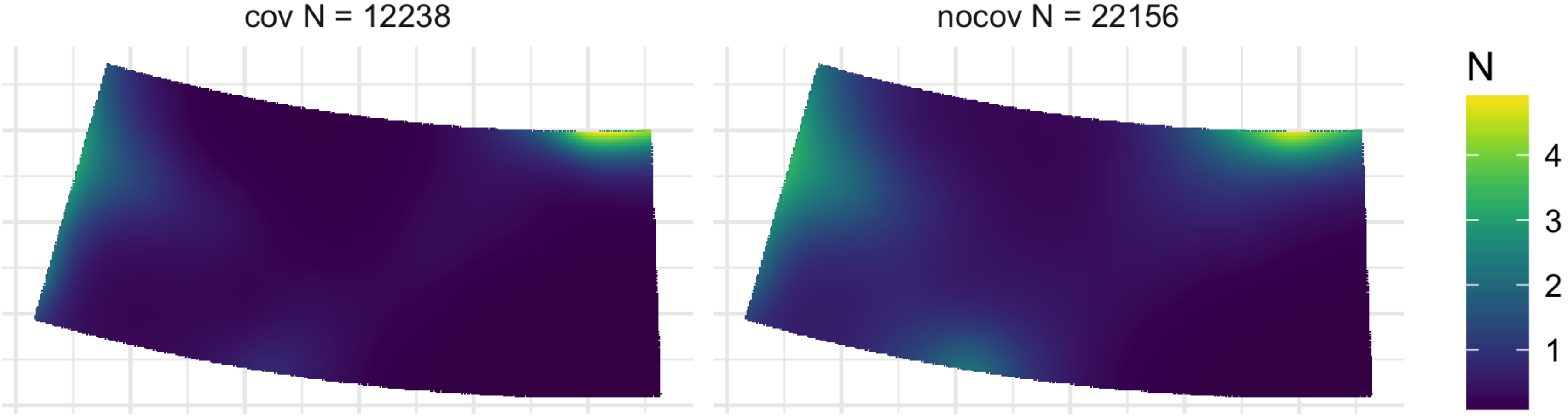
Visibility during POWER 2014



Thanks to Hiroto Murase and co. for this data!

Covariates can make a big difference!

- Same data, same spatial model
- With weather covariates and without



Disappointment

- Sometimes you don't have enough data
- Or, enough coverage
- Or, the right covariates

Sometimes, you can't build a spatial model

Segmenting

- Example on course site
- Length of $\approx 2w$ is reasonable
- Too big: no detail
- Too small: all 0/1
- See also Redfern et al., (2008)

Getting help

Resources

- Course reading list has pointers to these topics
- DenMod wiki with FAQ and more
- Distance sampling Google Group
  - Friendly, helpful, low traffic
  - see distancesampling.org/distancelist.html

That's all folks!