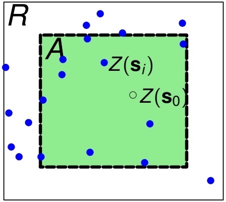
Estimating Abundance from Counts in Large Data Sets of Irregularly-Spaced Plots using Spatial Basis Functions

Jay Ver Hoef

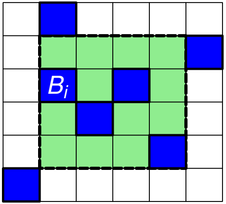
NOAA National Marine Mammal Lab
NOAA Fisheries
International Arctic Research Center
Fairbanks, Alaska, USA

Introduction

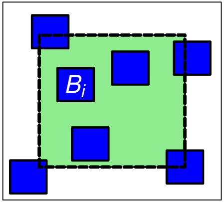
1) Block Kriging



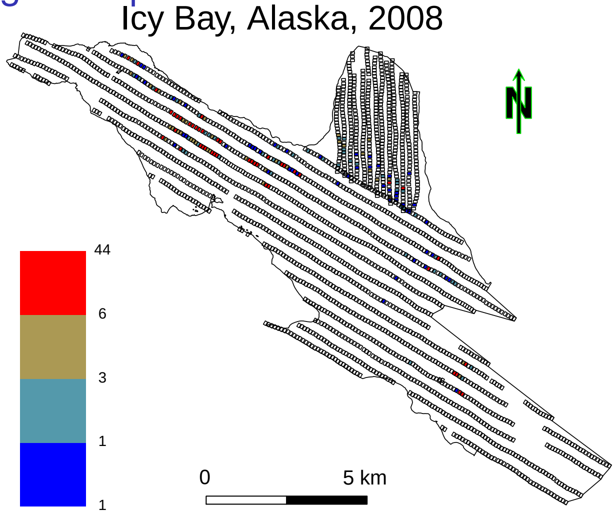
2)Block Prediction for Finite Populations on a Grid



3)Block Prediction for Finite Populations Irregularly Spaced



Motivating Example



Goals

An estimator that is:

- fast to compute, robust, and requires few modeling decisions, similar to classical survey methods,
- based only on counts within plots; actual spatial locations of animals are unknown,
- for the actual number of seals, not the mean of some assumed process that generated the data,
- have a variance estimator with a population correction factor that shrinks to zero as the proportion of the study area that gets sampled goes to one,
- unbiased with valid confidence intervals,
- able to accommodate nonstationary variance and excessive zeros throughout the area

Inhomogeneous Spatial Point Processes

$T(V)$ is the total number of points in planar region V

$$\lambda(\mathbf{s}) = \lim_{|dx|\to 0} \frac{E(T(dx))}{|dx|}$$

Expected abundance in $A \subseteq R$:

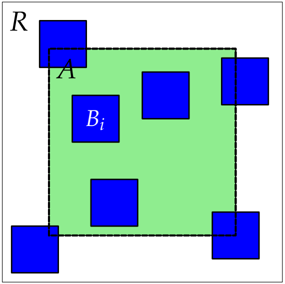
$$\mu(A) = \int_A \lambda(\mathbf{u}|\boldsymbol{\theta})d\mathbf{u}$$

Abundance is assumed random

$$T(A) \sim \text{Poi}(\mu(A))$$

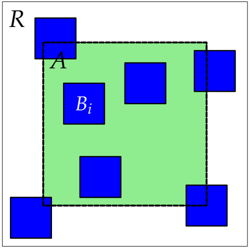
Resulting in an observed pattern $\mathcal{S}^+ = (\mathbf{s}_1, \ldots, \mathbf{s}_N)$

Outline of an Estimator



- $\mathcal{B} = \cup_{i=1}^{n}(B_i \cap A)$
- $\mathcal{U} \equiv \overline{\mathcal{B}} \cap A$
- $T(A) = T(\mathcal{B}) + T(\mathcal{U})$
- $T(\mathcal{U}) \sim \text{Poi}(\mu(\mathcal{U}))$
- $\mu(\mathcal{U}) = \int_{\mathcal{U}} \lambda(\mathbf{u}|\boldsymbol{\theta})d\mathbf{u}$
- $\widehat{T}(A) = T(\mathcal{B}) + \widehat{T}(\mathcal{U})$
- $T(\mathcal{B}) \to T(A) \Rightarrow \widehat{T}(A) \to T(A)$

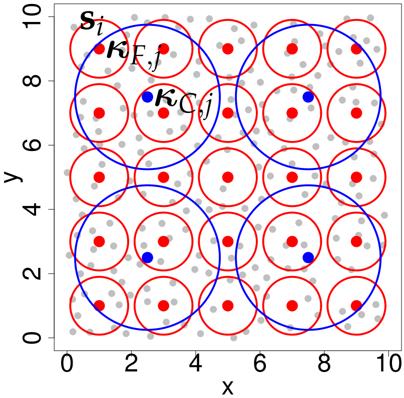
From IPP to Poisson Regression



- $Y(B_i) \sim \text{Poi}(\mu(B_i))$
- $\mu(B_i) = \int_{B_i} \lambda(\mathbf{u}|\boldsymbol{\theta})d\mathbf{u}$
- Let $\mathbf{s}_i$ be centroid of B_i
- $\mu(B_i) \approx |B_i|\lambda(\mathbf{s}_i|\boldsymbol{\theta})$
- $\log(\mu(B_i)) = \log(|B_i|) + \log(\lambda(\mathbf{s}_i|\boldsymbol{\theta}))$
- $\log(\lambda(\mathbf{s}_i|\boldsymbol{\theta})) = \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta}$

Now us spatial basis functions to generate $\mathbf{x}(\mathbf{s}_i)$

Spatial Basis Functions



- $C(h;\rho) = \exp(-h^2/\rho)$
- $\mathbf{X}_{i,j} = C(\|\mathbf{s}_i - \boldsymbol{\kappa}_{F,j})\|;\rho_F)$; $j = 2,\ldots,K_F+1$
- $\mathbf{X}_{i,j} = C(\|\mathbf{s}_i - \boldsymbol{\kappa}_{C,j})\|;\rho_C)$; $j = K_F+2,\ldots,K_F+K_C+1$

knot location: k-means clustering of dense grid of spatial coordinates

Fitting the Model

minimize minus the log-likelihood:

$$-\ell(\boldsymbol{\rho}, \boldsymbol{\beta}; \mathbf{y}) \propto \sum_{i=1}^{n} |B_i| \exp(\mathbf{x}_{\boldsymbol{\rho}}(\mathbf{s}_i)'\boldsymbol{\beta}) - y_i \log|B_i| - y_i \mathbf{x}_{\boldsymbol{\rho}}(\mathbf{s}_i)'\boldsymbol{\beta}$$

Two-part algorithm:

- Condition on ρ and use IWLS to estimate β (with offset for $|B_i|$, ala GLMs)
- optimize for ρ numerically

Back to the Estimator

- $\widehat{T}(A) = T(\mathcal{B}) + \widehat{T}(\mathcal{U})$
- $\widehat{T}(\mathcal{U}) = \mu(\mathcal{U}) = \int_{\mathcal{U}} \lambda(\mathbf{u}|\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\beta}}) d\mathbf{u}$
- $\lambda(\mathbf{u}|\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{\beta}}) = \exp(\mathbf{x}_{\hat{\boldsymbol{\rho}}}(\mathbf{u})'\hat{\boldsymbol{\beta}})$

Approximate integral with dense grid of n_p points within $\mathbf{u}_j \in \mathcal{U}$.

$$\widehat{T}(A) = T(\mathcal{B}) + \sum_{j=1}^{n_p} |U_i| \exp(\mathbf{x}_{\hat{\boldsymbol{\rho}}}(\mathbf{u}_j)'\hat{\boldsymbol{\beta}})$$

where $|U_i|$ is a small area around each $\mathbf{u}_j$

Variance

$\text{MSPE}(\hat{T}(A)) = E[(\hat{T}(A) - T(A))^2; \boldsymbol{\beta}] = E[(\hat{T}(\mathcal{U}) - T(\mathcal{U}))^2; \boldsymbol{\beta}]$
Note: as $\mathcal{U} \cap A \to \varnothing \Rightarrow \text{MSPE}(\widehat{T}(A)) \to 0$
From IPP assumption: $\hat{T}(\mathcal{U})$ independent from $T(\mathcal{U})$.
Assuming unbiasedness, $E[(\hat{T}(\mathcal{U})] = E[T(\mathcal{U})]$,

$$\begin{aligned} \text{MSPE} &= \text{var}[T(\mathcal{U}); \boldsymbol{\beta}] + \text{var}[\hat{T}(\mathcal{U}); \boldsymbol{\beta}] \\ &= \mu(\mathcal{U}; \boldsymbol{\beta}) + \text{var}[\hat{T}(\mathcal{U}); \boldsymbol{\beta}] \end{aligned}$$

Now, what about $\text{var}[\hat{T}(\mathcal{U}); \boldsymbol{\beta}]$?

Variance

Recall delta method result: $\mathrm{var}(f(\mathbf{y})) \approx \mathbf{d}'\boldsymbol{\Sigma}\mathbf{d}$

Jay M. Ver Hoef (2012) Who Invented the Delta Method? The American Statistician, 66:2, 124-127

where $\mathrm{var}(\mathbf{y}) = \boldsymbol{\Sigma}$ and $d_i = \partial f(\mathbf{y})/\partial y_i$

$$d_i = \frac{\partial \hat{T}(\mathcal{U})}{\partial \beta_i} = \int_{\mathcal{U}} x_i(\mathbf{u}) \exp(\mathbf{x}(\mathbf{u})'\hat{\boldsymbol{\beta}}) d\mathbf{u} \approx \frac{|\mathcal{U}|}{n_p} \sum_{i=1}^{n_p} x_i(\mathbf{s}_i) \exp(\mathbf{x}(\mathbf{s}_i)'\hat{\boldsymbol{\beta}})$$

From Rathbun and Cressie, (1994), if $\hat{\beta}$ is MLE,

$$\hat{\boldsymbol{\Sigma}} = \left[\sum_{i=1}^{n} \int_{B_i} \mathbf{x}(\mathbf{s})\mathbf{x}(\mathbf{s})' \exp(\mathbf{x}(\mathbf{s})'\hat{\boldsymbol{\beta}}) d\mathbf{s} \right]^{-1} \approx \left[|B| \sum_{i=1}^{n} \mathbf{x}(\mathbf{s}_i)\mathbf{x}(\mathbf{s}_i)' \exp(\mathbf{x}(\mathbf{s}_i)'\hat{\boldsymbol{\beta}}) \right]^{-1}$$

if $|B_i| = |B| \; \forall \; i$.

Rathbun, S. L. and Cressie, N. (1994), “Asymptotic Properties of Estimators for the Parameters of Spatial Inhomogeneous Poisson Point Processes,” Advances in Applied Probability, 26, 122-154.

Summary

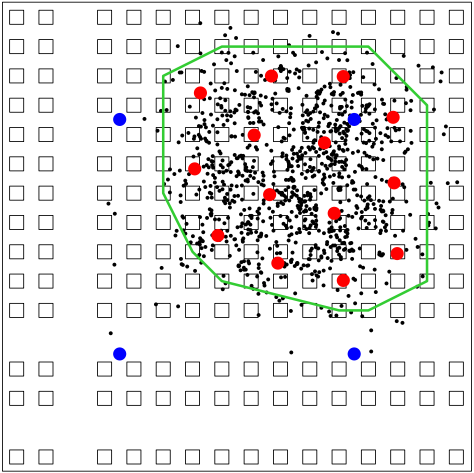
$$\widehat{T}(A) = T(\mathcal{B}) + \frac{|\mathcal{U}|}{n_p} \sum_{j=1}^{n_p} \exp(\mathbf{x}_{\hat{\boldsymbol{\rho}}}(\mathbf{u}_j)'\hat{\boldsymbol{\beta}})$$

$$\begin{aligned}\widetilde{\mathrm{var}}(\widehat{T}(A)) = \quad & \frac{|\mathcal{U}|}{n_p} \sum_{j=1}^{n_p} \exp(\mathbf{x}_{\hat{\boldsymbol{\rho}}}(\mathbf{u}_j)'\hat{\boldsymbol{\beta}}) + \\ & \mathbf{d}' \left[|B| \sum_{i=1}^{n} \mathbf{x}(\mathbf{s}_i)\mathbf{x}(\mathbf{s}_i)' \exp(\mathbf{x}(\mathbf{s}_i)'\hat{\boldsymbol{\beta}}) \right]^{-1} \mathbf{d}\end{aligned}$$

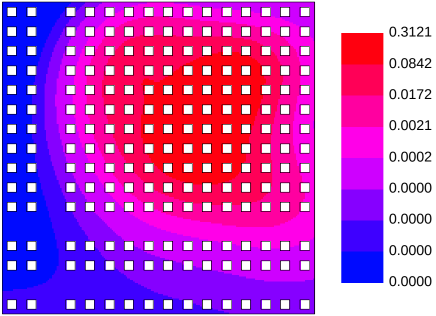
where

$$d_i = \frac{|\mathcal{U}|}{n_p} \sum_{i=1}^{n_p} x_i(\mathbf{s}_i) \exp(\mathbf{x}(\mathbf{s}_i)'\hat{\boldsymbol{\beta}})$$

Simulated Example

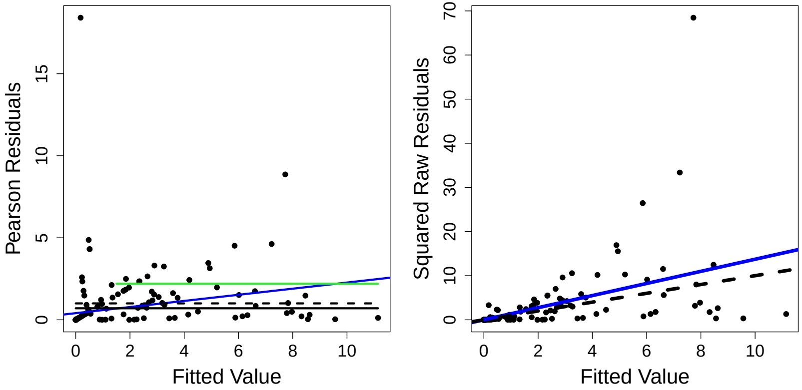


Simulated Example



True abundance was 1079
Estimated abundance was 1143 with standard error 62

Residuals Plots



Overdispersion Estimators

- The traditional estimator:

$$\omega_{OD} = \max\left(1, \frac{1}{n-r}\sum_{i=1}^{n}\frac{(y_i-\phi_i)^2}{\phi_i}\right)$$

  where r is the rank of $\mathbf{X}$.
- Weighted regression estimator:

$$\omega_{WR} = \max\left(1, \arg\min_{\omega}\sum_{i=1}^{n}\sqrt{\phi_i}[(y_i-\phi_i)^2-\omega\phi_i]^2\right),$$

  where $\sqrt{\phi_i}$ were the weights

Overdispersion Estimators

- Trimmed Mean:

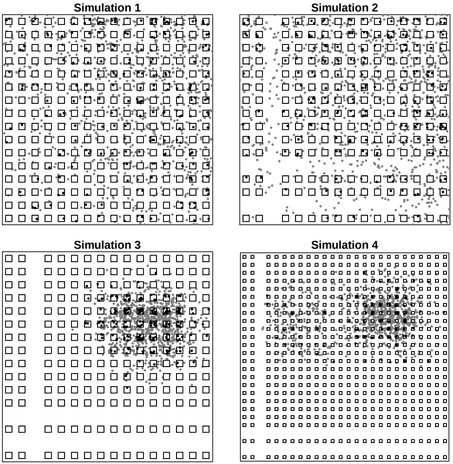
$$\omega_{TG}(p) = \max\left(1, \frac{1}{n - \lfloor np \rfloor - r} \sum_{i=\lfloor np \rfloor + 1}^{n} \frac{(y_{(i)} - \phi_{(i)})^2}{\phi_{(i)}}\right)$$

where $0 \le p \le 1$, $y_{(i)}$ and $\phi_{(i)}$ are ordered values, and $\lfloor x \rfloor$ rounds x down to the nearest integer.

Adjusted Variance Estimators

- $\widehat{\text{var}}_{OD}(\widehat{T}(A)) = \omega_{OD}\widetilde{\text{var}}(\widehat{T}(A))$
- $\widehat{\text{var}}_{WR}(\widehat{T}(A)) = \omega_{WR}\widetilde{\text{var}}(\widehat{T}(A))$
- $\widehat{\text{var}}_{TG}(\widehat{T}(A); p) = \omega_{TG}(p)\widetilde{\text{var}}(\widehat{T}(A))$
- $\widehat{\text{var}}_{TL}(\widehat{T}(A); p) = \frac{|\mathcal{U}|}{n_p} \sum_{j=1}^{n_p} \exp(\mathbf{x}_{\hat{\boldsymbol{\rho}}}(\mathbf{u}_j)'\hat{\boldsymbol{\beta}}) \times$ $\max(1, \omega_{TG}(p) I(\exp(\mathbf{x}(\mathbf{s}_j)'\hat{\boldsymbol{\beta}}) \geq \phi_{(\lfloor np \rfloor)}) +$ $((1-p)\omega_{TG}(p) + p) \times$ $\mathbf{d}' \left[|B| \sum_{i=1}^{n} \mathbf{x}(\mathbf{s}_i)\mathbf{x}(\mathbf{s}_i)' \exp(\mathbf{x}(\mathbf{s}_i)'\hat{\boldsymbol{\beta}}) \right]^{-1} \mathbf{d}$
  where $I(\cdot)$ is the indicator function

Simulations



Simulation Experiment 1

		Knots			
	SRS	$K_C = 3$ $K_F = 8$	$K_C = 4$ $K_F = 14$	$K_C = 5$ $K_F = 20$	$K_C = 6$ $K_F = 26$
bias	6.425	-1.277	-9.735	7.048	5.941
RMSPE	58.060	57.493	59.036	58.243	58.038
CI90	0.914	0.898	0.883	0.892	0.895
CI90$_{OD}$		0.918	0.927	0.897	0.904
CI90$_{WR}$		0.900	0.892	0.895	0.895
CI90$_{TG}$		0.914	0.920	0.939	0.957
CI90$_{TL}$		0.906	0.906	0.917	0.930
fail rate	0.000	0.000	0.000	0.000	0.000

Simulation Experiment 2

	SRS	Knots			
		$K_C = 3$ $K_F = 8$	$K_C = 4$ $K_F = 14$	$K_C = 5$ $K_F = 20$	$K_C = 6$ $K_F = 26$
bias	79.234	-1.333	-7.632	14.511	13.856
RMSPE	104.979	66.347	68.846	68.311	68.527
CI90	0.726	0.876	0.860	0.891	0.889
CI90$_{OD}$		0.908	0.928	0.901	0.902
CI90$_{WR}$		0.888	0.868	0.894	0.891
CI90$_{TG}$		0.902	0.903	0.952	0.980
CI90$_{TL}$		0.895	0.880	0.927	0.940
fail rate	0.000	0.000	0.000	0.000	0.001

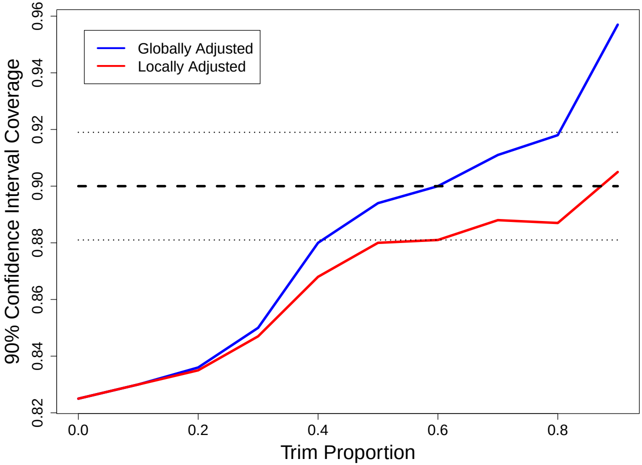
Simulation Experiment 3

		Knots			
	SRS	$K_C = 3$ $K_F = 8$	$K_C = 4$ $K_F = 14$	$K_C = 5$ $K_F = 20$	$K_C = 6$ $K_F = 26$
bias	214.816	-2.389	-4.365	-2.919	-1.637
RMSPE	235.713	79.207	79.250	79.285	80.175
CI90	0.774	0.775	0.772	0.781	0.777
CI90_{OD}		0.801	0.783	0.789	0.782
CI90_{WR}		0.918	0.906	0.865	0.837
CI90_{TG}		0.923	0.930	0.929	0.946
CI90_{TL}		0.871	0.883	0.878	0.903
fail rate	0.000	0.000	0.000	0.000	0.018

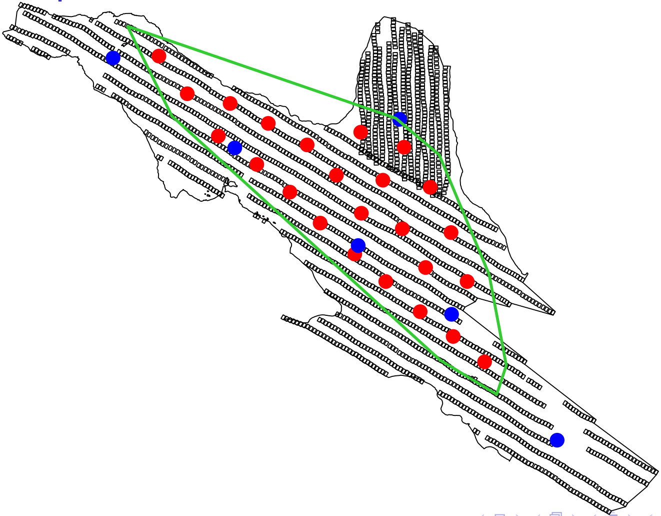
Simulation Experiment 4

		Knots			
	SRS	$K_C = 3$ $K_F = 8$	$K_C = 5$ $K_F = 16$	$K_C = 7$ $K_F = 24$	$K_C = 9$ $K_F = 32$
bias	148.523	5.179	3.440	7.287	14.629
RMSPE	163.516	60.403	61.021	62.102	64.136
CI90	0.834	0.831	0.825	0.826	0.833
CI90_{OD}		0.844	0.831	0.827	0.833
CI90_{WR}		0.939	0.928	0.923	0.913
CI90_{TG}		0.929	0.919	0.907	0.920
CI90_{TL}		0.893	0.892	0.884	0.886
fail rate	0.000	0.000	0.000	0.000	0.242

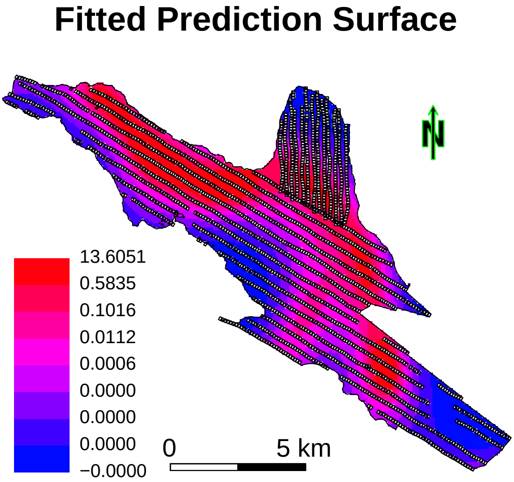
Effect of p in Trimmed Overdispersion Estimator



Real Example



Real Example



Real Example

```
sealDensity <- sum(plots@data[, "counts"])/(sCSout$propSurveyed * totalArea)
sealDensity * totalArea


## [1] 3960


summary(sCSout)


##
## Estimates:
##
## Total:
## [1] 4068
## Standard Errors:
##      SE SE.ODTrad SE.ODTrimGlobal SE.ODTrimLocal SE.ODRegr
## 1 113.7      1379             236          198.5     403.1
##
##
## Range Parameters:
##   coarseScale fineScale
## 1        9516      2974
##
##
## Proportion Surveyed:
## [1] 0.253
```

Recall the Goals

An estimator that is:

- fast to compute, robust, and requires few modeling decisions, similar to classical survey methods,
- based only on counts within plots; actual spatial locations of animals are unknown,
- for the actual number of seals, not the mean of some assumed process that generated the data,
- have a variance estimator with a population correction factor that shrinks to zero as the proportion of the study area that gets sampled goes to one,
- unbiased with valid confidence intervals,
- able to accommodate nonstationary variance and excessive zeros throughout the area