Block Prediction for a Finite Grid

Jay Ver Hoef

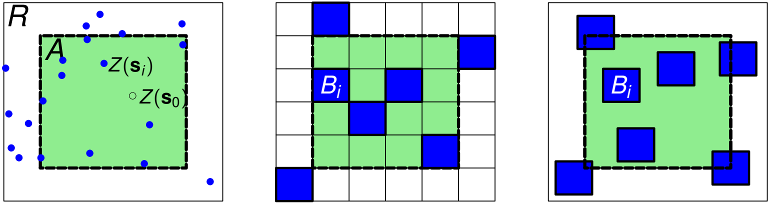
National Marine Mammal Lab
NOAA Fisheries
International Arctic Research Center
Fairbanks, Alaska, USA

Introduction

1) Block Kriging

2)Block Prediction for Finite Populations on a Grid

3)Block Prediction for Finite Populations Irregularly Spaced



Ver Hoef, J.M. 2001. Predicting finite populations from spatially correlated data. *2000 Proceedings of the Section on Statistics and the Environment of the American Statistical Association*, pgs. 93 - 98.
Ver Hoef, J.M. 2008. Spatial methods for plot-based sampling of wildlife populations. *Environmental and Ecological Statistics* **15**: 3 - 13.

Review of BLUP and Point Prediction

$$\begin{pmatrix} \mathbf{z} \\ Z(\mathbf{s}_0) \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \mathbf{x}(\mathbf{s}_0)' \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\epsilon} \\ \epsilon(\mathbf{s}_0) \end{pmatrix}$$

$$\mathrm{cov} \begin{pmatrix} \boldsymbol{\epsilon} \\ \epsilon(\mathbf{s}_0) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{c} \\ \mathbf{c}' & \sigma_0^2 \end{pmatrix}$$

Best Linear Unbiased Prediction (BLUP) (or [Universal] Kriging) minimize: $E(\boldsymbol{\lambda}'\mathbf{z} - Z(\mathbf{s}_0))^2$ subject to $E[\boldsymbol{\lambda}'\mathbf{z}] = E[Z(\mathbf{s}_0)] \ \forall \ \boldsymbol{\beta}$.
Unbiasedness $\Rightarrow \mathbf{X}'\boldsymbol{\lambda} = \mathbf{x}(\mathbf{s}_0)$
Minimization yields the solution to:

$$\begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{X} \\ \mathbf{X}' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda} \\ \mathbf{m} \end{pmatrix} = \begin{pmatrix} \mathbf{c} \\ \mathbf{x}(\mathbf{s}_0) \end{pmatrix}$$

BLUP for Finite Populations

$$\begin{pmatrix} \mathbf{z}_s \\ \mathbf{z}_u \end{pmatrix} = \begin{pmatrix} \mathbf{X}_s \\ \mathbf{X}_u \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\epsilon}_s \\ \boldsymbol{\epsilon}_u \end{pmatrix}$$

$$\operatorname{cov}\begin{pmatrix} \boldsymbol{\epsilon}_s \\ \boldsymbol{\epsilon}_u \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{s,s} & \boldsymbol{\Sigma}_{s,u} \\ \boldsymbol{\Sigma}'_{s,u} & \boldsymbol{\Sigma}_{s,s} \end{pmatrix}$$

Best Linear Unbiased Prediction (BLUP) (or [Universal] Kriging)
minimize: $E(\boldsymbol{\lambda}'\mathbf{z}_s - \mathbf{b}'\mathbf{z})^2$ subject to $E[\boldsymbol{\lambda}'\mathbf{z}_s] = E[\mathbf{b}'\mathbf{z}] \;\forall\; \boldsymbol{\beta}$.
Unbiasedness $\Rightarrow \mathbf{X}_s'\boldsymbol{\lambda} = \mathbf{X}_s'\mathbf{b}_s + \mathbf{X}_u'\mathbf{b}_u$
Minimization yields the solution to:

$$\begin{pmatrix} \boldsymbol{\Sigma}_{s,s} & \mathbf{X}_s \\ \mathbf{X}_s' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda} \\ \mathbf{m} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{s,s} & \boldsymbol{\Sigma}_{u,u} \\ \mathbf{X}_s' & \mathbf{X}_u' \end{pmatrix} \begin{pmatrix} \mathbf{b}_s \\ \mathbf{b}_u \end{pmatrix}$$

BLUP for Finite Populations

$$\boldsymbol{\lambda}'\mathbf{z} = \mathbf{b}_s'\mathbf{z}_s + \mathbf{b}_u\hat{\mathbf{z}}_u$$

where

$$\hat{\mathbf{z}}_u = \boldsymbol{\Sigma}_{u,s}\boldsymbol{\Sigma}_{s,s}^{-1}(\mathbf{z}_s - \hat{\boldsymbol{\mu}}_s) + \hat{\boldsymbol{\mu}}_u,$$

$$\hat{\boldsymbol{\mu}}_u = \mathbf{X}_u\hat{\boldsymbol{\beta}}_{GLS}, \quad \hat{\boldsymbol{\mu}}_s = \mathbf{X}_s\hat{\boldsymbol{\beta}}_{GLS}, \quad \hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}_s'\boldsymbol{\Sigma}_{s,s}^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s'\boldsymbol{\Sigma}_{s,s}^{-1}\mathbf{z}_s$$

and

$$E(\boldsymbol{\lambda}'\mathbf{z}_s - \mathbf{b}\mathbf{z})^2 = \mathbf{b}'\boldsymbol{\Sigma}\mathbf{b} - \mathbf{c}_\mathbf{b}'\boldsymbol{\Sigma}_{s,s}^{-1}\mathbf{c}_\mathbf{b} + \mathbf{d}_\mathbf{b}'(\mathbf{X}_s'\boldsymbol{\Sigma}_{s,s}^{-1}\mathbf{X}_s)^{-1}\mathbf{d}_\mathbf{b}$$

where

$$\mathbf{c}_\mathbf{b} = \boldsymbol{\Sigma}_{s,s}\mathbf{b}_s + \boldsymbol{\Sigma}_{s,u}\mathbf{b}_u, \quad \mathbf{d}_\mathbf{b} = \mathbf{X}'\mathbf{b} - \mathbf{X}_s'\boldsymbol{\Sigma}_{s,s}^{-1}\mathbf{c}_\mathbf{b}$$

Connections to Sampling Theory

Let
$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_n \\ \mathbf{1}_{N-n} \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2\mathbf{I}_n & \mathbf{0} \\ \mathbf{0}' & \sigma^2\mathbf{I}_{N-n} \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} \mathbf{1}_n(1/N) \\ \mathbf{1}_{N-n}(1/N) \end{pmatrix}$$
Then

$$\boldsymbol{\lambda}'\mathbf{z}_s = \bar{z}$$

and

$$E(\boldsymbol{\lambda}'\mathbf{z}_s - \mathbf{b}\mathbf{z})^2 = \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right)$$

Connections to Sampling Theory

Let

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} \\ \mathbf{1}_{N_1-n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{N_2-n_2} \end{pmatrix} \quad \boldsymbol{\Sigma}_{s,s} = \begin{pmatrix} \sigma_1^2\mathbf{I}_{n_1} & \mathbf{0} \\ \mathbf{0}' & \sigma_2^2\mathbf{I}_{n_2} \end{pmatrix} \quad \boldsymbol{\Sigma}_{u,u} = \begin{pmatrix} \sigma_1^2\mathbf{I}_{N_1-n_1} & \mathbf{0} \\ \mathbf{0}' & \sigma_2^2\mathbf{I}_{N_2-n_2} \end{pmatrix}$$

and

$$\boldsymbol{\Sigma}_{s,u} = \mathbf{0}, \quad \mathbf{z} = (\mathbf{z}'_{s,1}, \mathbf{z}'_{s,2}, \mathbf{z}'_{u,1}, \mathbf{z}_{u,2})', \quad \mathbf{b} = \mathbf{1}_{N_1+N_2}$$

Then

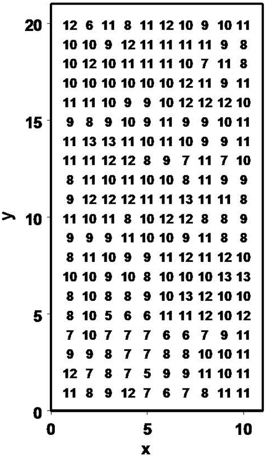
$$\boldsymbol{\lambda}'\mathbf{z}_s = N_1\bar{z}_{s,1} + N_2\bar{z}_{s,2}$$

and

$$E(\boldsymbol{\lambda}'\mathbf{z}_s - \mathbf{b}\mathbf{z})^2 = N_1^2\frac{\sigma_1^2}{n_1}\left(1 - \frac{n_1}{N_1}\right) + N_2^2\frac{\sigma_2^2}{n_2}\left(1 - \frac{n_2}{N_2}\right)$$

Simulations

Number of plant species in 70 cm $\times$ 70 cm plot



1000 samplings

$N = 200$

$n = 100$

Fixed Pattern, Random Samples

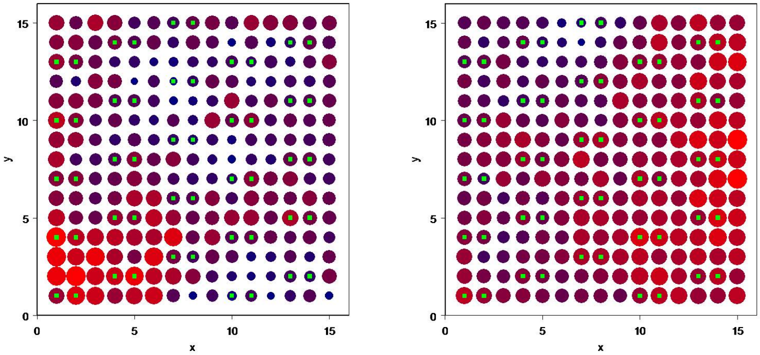
1000 random samples of size 100

Ver Hoef, J.M. 2002. Sampling and geostatistics for spatial data. *Ecoscience* **9**: 152 - 161.

Validation Statistics	Simple Random Sample	Finite Block Pred
Bias	-0.002	-0.001
RMSPE	0.121	0.106
RAEV	0.122	0.105
80%CI	0.802	0.806

Random Pattern, Fixed Samples

Ver Hoef, J.M. 2002. Sampling and geostatistics for spatial data. *Ecoscience* **9**: 152 - 161.



Random Pattern, Fixed Samples

Validation Statistics	SRS_{random}	$FPBK_{random}$	$FPBK_{fixed}$
Bias	0.522	-0.181	0.127
RMSPE	28.0	20.7	17.3
RAEV	28.0	20.3	17.5
80%CI	0.801	0.791	0.796

Simulation With Covariates

Ver Hoef, J.M. and Temesgen, H. 2013. A comparsion of the spatial linear model to nearest neighbor (k-NN) methods for forestry applications. *PloS ONE* **8(3)**: e59129. doi:10.1371/journal.pone.0059129.

$$\begin{aligned} \mathbf{w}_1 &= \mathbf{z}_1 + \boldsymbol{\epsilon}_1 \\ \mathbf{w}_\eta &= \phi_{\eta-1}\mathbf{w}_{\eta-1} + \mathbf{z}_\eta + \boldsymbol{\epsilon}_\eta \\ \mathbf{x}_\eta &= \boldsymbol{\mu}_\eta + \mathbf{w}_\eta \\ \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{z}_y + \boldsymbol{\epsilon}_y \end{aligned}$$

$$\mathbf{V}[j,j';\sigma^2,\rho] = \sigma^2\left(1 - \frac{3}{2}\frac{d_{j,j'}}{\rho} + \frac{1}{2}\frac{d_{j,j'}}{\rho^3}\right) I\left(\frac{d_{j,j'}}{\rho} \leq 1\right)$$

8 spatially-patterned and cross-correlated covariates
2 excluded, 2 with $\beta_i = 0$ included

Prediction Methods

- mah1 k-NN that uses Mahalanobis distance with $k = 1$.
- mah5: k-NN that uses Mahalanobis distance with $k = 5$.
- msn1: k-NN that uses most significant neighbor (MSN) with $k = 1$.
- msn5: k-NN that uses MSN with $k = 1$.
- best: k-NN that uses both Mahalanobis distance and MSN, and tries $k = 1, 2, \ldots, 30$, and then chooses the distance matrix and k with the smallest cross-validation RMSPE from the observed data.
- slm: a spatial linear model using the same covariates as all k-NN methods as main effects only, with an exponential autocovariance model estimated by REML, and using FPBK prediction and variance equations.
- lm: multiple regression like slmMain but assuming all random errors are independent.

Performance Measures

- Root-mean-squared-prediction error (RMSPE):

$$\text{RMSPE} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}(\hat{\theta}_j - \theta_j)^2},$$

- SRB: signed relative bias.

$$\text{SRB} = \text{sign}(\tau)\sqrt{\frac{\tau^2}{\text{MSPE} - \tau^2}},$$

where

$$\tau = \frac{1}{m}\sum_{j=1}^{m}(\hat{\theta}_j - \theta_j),$$

and sign(τ) is the sign (positive or negative) of τ.

- PIC90: 90% prediction interval coverage

$$\text{PIC90} = \frac{1}{m}\sum_{j=1}^{m} I\left(\left(\hat{\theta}_j - 1.645\hat{\text{se}}(\hat{\theta}_j)\right) < \theta_j \ \&\ \theta_j < \left(\hat{\theta}_j + 1.645\hat{\text{se}}(\hat{\theta}_j)\right)\right),$$

where $\hat{\text{se}}(\hat{\theta}_j)$ is the estimated standard error of $\hat{\theta}_j$.

Gaussian Simulations

Table: Performance summaries from 2000 simulated spatial Gaussian data sets with 100 samples and 300 predictions each. Prediction methods form the columns and are described in Section 3.4. Performance measures form the rows and are described in Section 3.3.

	mah1	mah5	msn1	msn5	best	lm	slm
			Point				
RMSPE[a]	9.329	7.451	5.379	4.423	4.456	3.892	2.443
SRB[b]	-0.006	-0.009	0	-0.004	-0.004	-0.002	0.001
PIC90[c]	0.897	0.9	0.887	0.889	0.88	0.896	0.892
			Total				
RMSPE	262.6	289.8	174.3	153.3	154.5	139.3	87.8
SRB	-0.058	-0.067	-0.003	-0.034	-0.034	-0.02	0.009
PIC90	0.952	0.87	0.914	0.886	0.874	0.88	0.887

[a] Root-mean-squared-prediction error
[b] Signed relative bias
[c] 90% prediction interval coverage

Poisson Simulations

Table: Performance summaries from 2000 simulated Poisson data sets with 100 samples and 300 predictions each. Prediction methods form the columns and are described in Section 3.4. Performance measures form the rows and are described in Section 3.3.

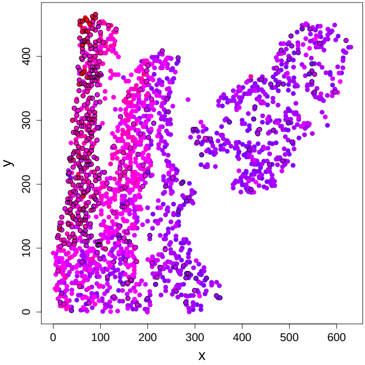
	mah1	mah5	msn1	msn5	best	lm	slm
				Total			
RMSPE	320	295.9	296.3	262.3	272.2	283.1	226.1
SRB	-0.137	-0.188	-0.047	-0.135	-0.182	-0.033	-0.005
PIC90	0.912	0.842	0.9	0.867	0.83	0.86	0.858

Bernoulli Simulations

Table: Performance summaries from 2000 simulated binary data sets with 100 samples and 300 predictions each. Prediction methods form the columns and are described in Section 3.4. Performance measures form the rows and are described in Section 3.3.

	mah1	mah5	msn1	msn5	best	lm	slm
				Proportion			
RMSPE	0.0395	0.0394	0.0387	0.0334	0.0343	0.0329	0.0298
SRB	0.072	0.09	0.003	0.019	0.079	0.018	0.014
PIC90	0.919	0.841	0.913	0.882	0.84	0.886	0.884

Forestry Data Resampling PMAI



- response: maximum potential mean annual increment (PMAI)
- covariates: 1) temperature, 2) precipitation, 3)Climate Moisture Index, 4) an indicator variable for shade tolerance based on Western Hemlock trees, and 5) elevation

PMAI Resampling

Table: Performance summaries for 500 resamplings of PMAI forest data with 386 samples and 1500 predictions each. Prediction methods form the columns and are described in Section 3.4. Performance measures form the rows and are described in Section 3.3.

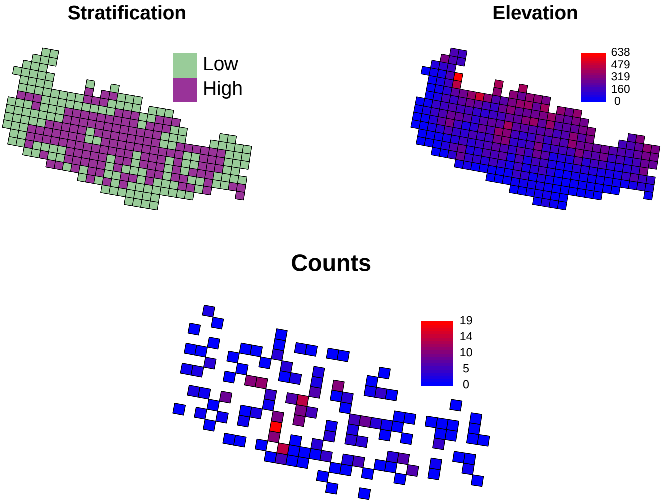
	mah1	mah5	msn1	msn5	best	lm	slm
				Total			
RMSPE	219.1	230.7	243.3	200.9	223.2	197	180.4
SRB	0.437	0.712	0.064	-0.019	0.446	0.082	0.058
PIC90	0.944	0.838	0.948	0.922	0.834	0.904	0.904

PMAI Resampling with Unbalanced Sampling

Table: Performance summaries for 500 resamplings of PMAI forest data with 386 spatially unbalanced samples and 1500 predictions each.

	mah1	mah5	msn1	msn5	best	lm	slm
				Total			
RMSPE	637.9	853.6	457.1	442.2	576.1	608.1	269
SRB	2.635	4.055	1.418	1.77	1.651	2.86	0.369
PIC90	0.248	0.010	0.626	0.438	0.308	0.128	0.92

Nome Moose Survey



Nome Moose Survey

```
library(spPlotSampCourse)
library(maptools)
path <- system.file("rawdata/moose", package = "spPlotSampCourse")
samplesFile <- paste(path, "/", "Samples", sep = "")
samples <- readShapePoly(samplesFile)
samples@data[, "x"] <- LLtoUTM(samples@data[, "CENTRLAT"], samples@data[, "CENTRLON"])[,
    "x"]
samples@data[, "y"] <- LLtoUTM(samples@data[, "CENTRLAT"], samples@data[, "CENTRLON"])[,
    "y"]
samples@data[, "TOTAL"] <- as.numeric(as.character(samples@data[, "TOTAL"]))
samples@data[, "b"] <- rep(1, times = length(samples@data[, 1]))
sdata <- samples@data
coordinates(sdata) <- c("x", "y")
```

Nome Moose Survey

```
moFit <- splmm(TOTAL ~ ELEVMEAN + STRAT, spdata = sdata, varComps = "exponential")
# summary(moFit)
summary(moFit)$coefficients

##                 Estimate Std. Error t value   Pr(>|t|)
## (Intercept)  2.592622   0.711248  3.6452 0.0004013
## ELEVMEAN     0.001019   0.003291  0.3097 0.7573331
## STRATL      -2.232373   0.611141 -3.6528 0.0003908

summary(moFit)$covparms

##   Variance Component Parameter Type  Estimate
## 1             nugget         nugget 1.928e-05
## 2        exponential         parsil 1.006e+01
## 3        exponential          range 1.043e+01
```

Nome Moose Survey

```
moFit <- splmm(TOTAL ~ STRAT - 1, spdata = sdata, varComps = "exponential")
summary(moFit)$coefficients
```

```
##        Estimate Std. Error t value  Pr(>|t|)
## STRATH   2.7442     0.5429  5.0547 1.609e-06
## STRATL   0.5289     0.5725  0.9239 3.574e-01
```

```
summary(moFit)$covparms
```

```
##   Variance Component Parameter Type  Estimate
## 1             nugget         nugget 1.287e-04
## 2        exponential         parsil 9.925e+00
## 3        exponential          range 1.023e+01
```

```
FPBK <- predictBlockFinPop(moFit, "b")
FPBK
```

```
##    Resp Wts  Pred Pred SE
## 1 TOTAL   b 554.6   60.62
```

Nome Moose Survey

```
samples.H <- samples[samples@data[, "STRAT"] == "H", ]
samples.L <- samples[samples@data[, "STRAT"] == "L", ]
sdataH <- samples.H@data
coordinates(sdataH) <- c("x", "y")
sdataL <- samples.L@data
coordinates(sdataL) <- c("x", "y")
moFitH <- splmm(TOTAL ~ 1, spdata = sdataH, varComps = "exponential")
summary(moFitH)$coefficients


##             Estimate Std. Error t value   Pr(>|t|)
## (Intercept)     2.78     0.7452   3.731 0.0003782


moFitL <- splmm(TOTAL ~ 1, spdata = sdataL, varComps = "exponential")
summary(moFitL)$coefficients


##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept)   0.3082      45.94 0.006709   0.9947
```

Nome Moose Survey

```
FPBK.H <- predictBlockFinPop(moFitH, "b")
FPBK.L <- predictBlockFinPop(moFitL, "b")
rbind(FPBK.H, FPBK.L)
```

```
##    Resp Wts   Pred Pred SE
## 1 TOTAL   b 501.26   38.40
## 2 TOTAL   b  51.38   14.08
```

```
FPBK.strat <- data.frame(Pred = FPBK.H[, "Pred"] + FPBK.L[, "Pred"], PredSE = sqrt(FPBK.H[,
    "Pred SE"]^2 + FPBK.L[, "Pred SE"]^2))
FPBK.strat
```

```
##    Pred PredSE
## 1 552.6   40.9
```

```
FPBK
```

```
##    Resp Wts  Pred Pred SE
## 1 TOTAL   b 554.6   60.62
```

Robustness of Block Prediction Methods

- BLUP is non-parametric
- REML are estimating equations, so covariance estimation is non-parametric
- Predictions are robust to mis-specification of covariance model
- Both predictor and predictand are linear combinations, so we can appeal to a correlated version of central limit theorem and use normal-distribution for probability statements (e.g., prediction intervals).