Spatial Modeling of Zero-Inflated Data with Copula Models

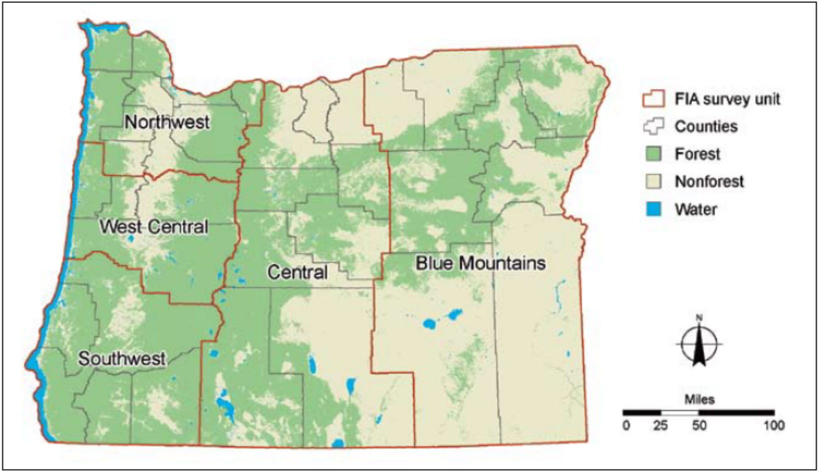
Lisa Madsen [1] Vicente Monleon [2]

[1] Oregon State University

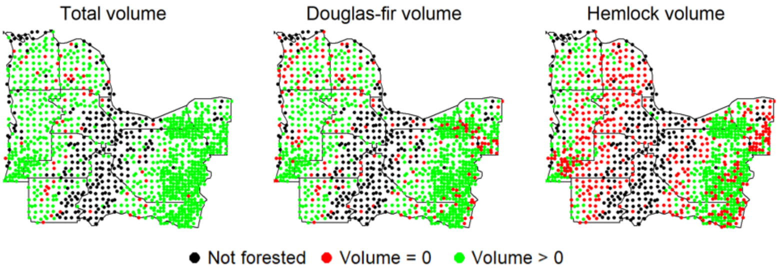
[2] USDA Forest Service, PNW Research Station

WNAR 2021

Forest Inventory and Analysis (FIA)



Variables of Interest



24% 0's　　34% 0's　　56% 0's

Challenges

- Spatial dependence
- Non-normality
- Zero-inflation

Challenges

- Spatial dependence
- Non-normality
- Zero-inflation
- Big data

Outline

Outline

Copula Models

Definition: A **copula** is a multivariate distribution with uniform marginals.

Copula Models

Definition: A **copula** is a multivariate distribution with uniform marginals.

Copula models are useful for modeling multivariate data with arbitrary marginal distributions.

Copula Models

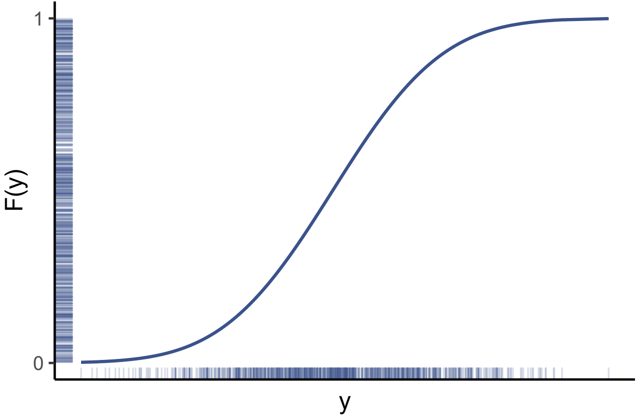
Definition: A **copula** is a multivariate distribution with uniform marginals.

Copula models are useful for modeling multivariate data with arbitrary marginal distributions.

The **Gaussian copula** allows us to adapt methodology based on the multivariate normal distribution to non-normal data.

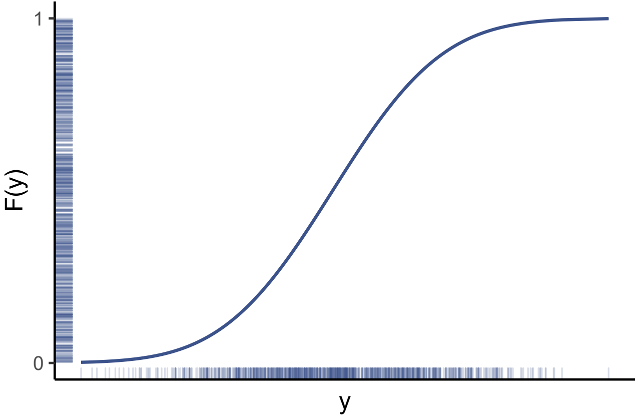
Copula Construction

If Y has continuous cdf $F(y)$, then $F(Y) \sim U(0, 1)$.



Copula Construction

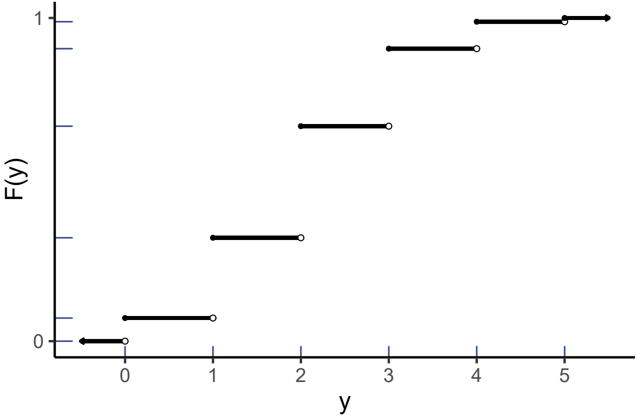
If Y has continuous cdf $F(y)$, then $F(Y) \sim U(0,1)$.



Conversely, if $U \sim U(0,1)$, then $F^{-1}(U) \sim F$.

Discrete Marginals

If Y is discrete, then $F^{-1}(U) \sim F$, but $F(Y)$ is not uniform.



Gaussian Copula

Notation:

- $\Phi_{\mathbf{\Sigma}}$ denotes the n-dimensional standard normal cdf with correlation matrix $\mathbf{\Sigma}$.

Gaussian Copula

Notation:

$\Phi_{\boldsymbol{\Sigma}}$ denotes the *n*-dimensional standard normal cdf with correlation matrix $\boldsymbol{\Sigma}$.

Φ denotes the univariate standard normal cdf.

Gaussian Copula

Notation:

$\Phi_{\boldsymbol{\Sigma}}$ denotes the n-dimensional standard normal cdf with correlation matrix $\boldsymbol{\Sigma}$.

Φ denotes the univariate standard normal cdf.

$F_1, \ldots, F_n$ are (preferably continuous) marginal cdfs.

Gaussian Copula

Notation:

$\Phi_{\boldsymbol{\Sigma}}$ denotes the *n*-dimensional standard normal cdf with correlation matrix $\boldsymbol{\Sigma}$.

Φ denotes the univariate standard normal cdf.

$F_1, \ldots, F_n$ are (preferably continuous) marginal cdfs.

Copula CDF:

$$C(\boldsymbol{y}; \boldsymbol{\Sigma}) = \Phi_{\boldsymbol{\Sigma}}[\Phi^{-1}\{F_1(y_1)\}, \ldots, \Phi^{-1}\{F_n(y_n)\}]$$

Gaussian Copula

Notation:

$\Phi_{\boldsymbol{\Sigma}}$ denotes the *n*-dimensional standard normal cdf with correlation matrix $\boldsymbol{\Sigma}$.

Φ denotes the univariate standard normal cdf.

$F_1, \ldots, F_n$ are (preferably continuous) marginal cdfs.

Copula CDF:

$$C(\boldsymbol{y}; \boldsymbol{\Sigma}) = \Phi_{\boldsymbol{\Sigma}}[\Phi^{-1}\{F_1(y_1)\}, \ldots, \Phi^{-1}\{F_n(y_n)\}]$$

Gaussian copula's normalizing transformation for continuous F_i:

$Y_i \sim F_i$

Gaussian Copula

Notation:

$\Phi_{\mathbf{\Sigma}}$ denotes the n-dimensional standard normal cdf with correlation matrix $\mathbf{\Sigma}$.

Φ denotes the univariate standard normal cdf.

$F_1, \ldots, F_n$ are (preferably continuous) marginal cdfs.

Copula CDF:

$$C(\boldsymbol{y}; \mathbf{\Sigma}) = \Phi_{\mathbf{\Sigma}}[\Phi^{-1}\{F_1(y_1)\}, \ldots, \Phi^{-1}\{F_n(y_n)\}]$$

Gaussian copula's normalizing transformation for continuous F_i:

$Y_i \sim F_i \;\Rightarrow\; F_i(Y_i) \sim \text{Uniform}(0, 1)$

Gaussian Copula

Notation:

$\Phi_{\boldsymbol{\Sigma}}$ denotes the *n*-dimensional standard normal cdf with correlation matrix $\boldsymbol{\Sigma}$.

Φ denotes the univariate standard normal cdf.

$F_1, \ldots, F_n$ are (preferably continuous) marginal cdfs.

Copula CDF:

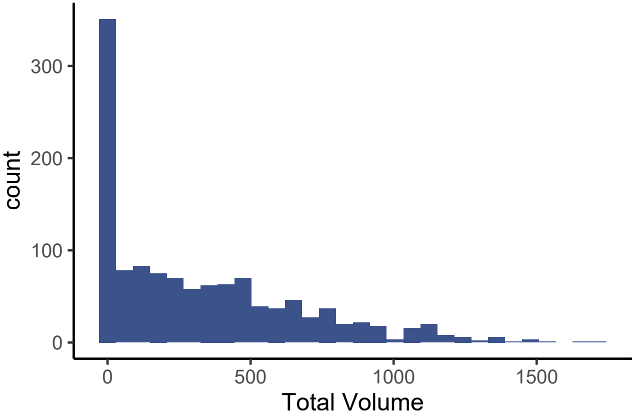
$$C(\boldsymbol{y}; \boldsymbol{\Sigma}) = \Phi_{\boldsymbol{\Sigma}}[\Phi^{-1}\{F_1(y_1)\}, \ldots, \Phi^{-1}\{F_n(y_n)\}]$$

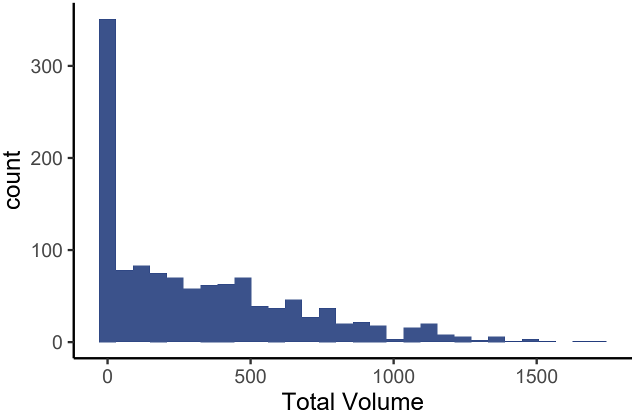
Gaussian copula's normalizing transformation for continuous F_i:

$$Y_i \sim F_i \;\Rightarrow\; F_i(Y_i) \sim \text{Uniform}(0,1) \;\Rightarrow\; \Phi^{-1}\{F_i(Y_i)\} \sim N(0,1)$$

Outline

Marginal Distributions





24% of plots have zero total volume.

Zero-inflated Lognormal Distribution

$$B \sim \text{Bernoulli}(1 - p)$$

Zero-inflated Lognormal Distribution

$$
\begin{aligned}
B &\sim \text{Bernoulli}(1-p) \\
W &\sim \text{Lognormal}(\mu, \sigma^2)
\end{aligned}
$$

Zero-inflated Lognormal Distribution

$$
\begin{aligned}
B &\sim \text{Bernoulli}(1-p) \\
W &\sim \text{Lognormal}(\mu, \sigma^2) \\
Y &= \begin{cases} 0 & B = 1 \\ W & B = 0 \end{cases}
\end{aligned}
$$

Zero-inflated Lognormal Distribution

$$
\begin{aligned}
B &\sim \text{Bernoulli}(1-p) \\
W &\sim \text{Lognormal}(\mu, \sigma^2) \\
Y &= \begin{cases} 0 & B = 1 \\ W & B = 0 \end{cases}
\end{aligned}
$$

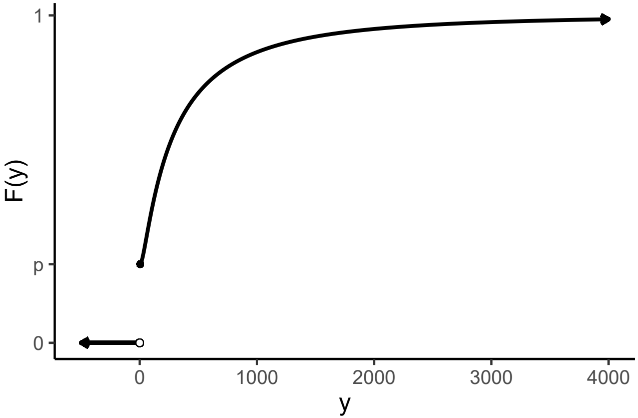
$$
P(Y=0) = p
$$

Zero-inflated Lognormal Distribution

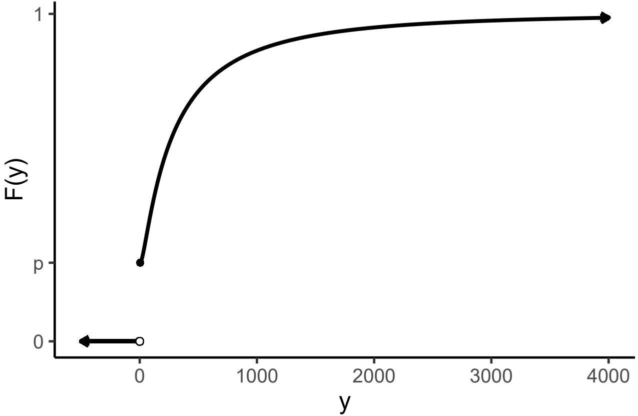
$$
\begin{aligned}
B &\sim \text{Bernoulli}(1-p) \\
W &\sim \text{Lognormal}(\mu, \sigma^2) \\
Y &= \begin{cases} 0 & B = 1 \\ W & B = 0 \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
P(Y=0) &= p \\
\log(W) &\sim N(\mu, \sigma^2)
\end{aligned}
$$

Zero-inflated Lognormal CDF



Zero-inflated Lognormal CDF



CDF discontinuous at 0.

Continuous Zero-inflated Lognormal Distribution

$$B \sim \text{Bernoulli}(1-p)$$

Continuous Zero-inflated Lognormal Distribution

$$
\begin{aligned}
B &\sim \text{Bernoulli}(1 - p) \\
W - \epsilon &\sim \text{Lognormal}(\mu, \sigma^2)
\end{aligned}
$$

Continuous Zero-inflated Lognormal Distribution

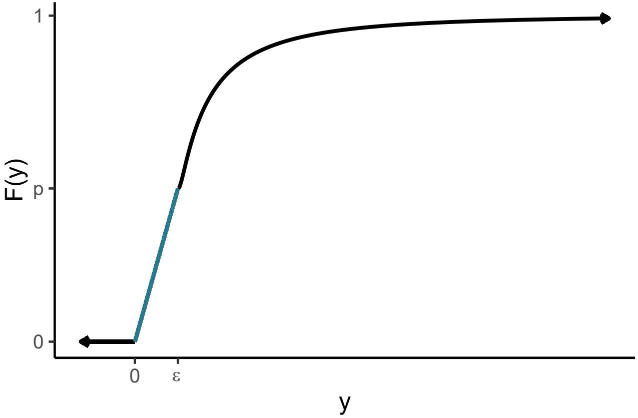
$$
\begin{aligned}
B &\sim \text{Bernoulli}(1-p) \\
W-\epsilon &\sim \text{Lognormal}(\mu, \sigma^2) \\
Y &= \begin{cases} \text{Uniform}(0,\epsilon) & B = 1 \\ W & B = 0 \end{cases}
\end{aligned}
$$

Continuous Zero-inflated Lognormal Distribution

$$
\begin{aligned}
B &\sim \text{Bernoulli}(1-p) \\
W-\epsilon &\sim \text{Lognormal}(\mu,\sigma^2) \\
Y &= \begin{cases} \text{Uniform}(0,\epsilon) & B=1 \\ W & B=0 \end{cases}
\end{aligned}
$$

$$
F(y) = \begin{cases} 0, & y < 0 \\ y \cdot p/\epsilon, & 0 \le y < \epsilon \\ p + (1-p)F_{\text{lnorm}}(y-\epsilon;\mu,\sigma^2), & y \ge \epsilon \end{cases}
$$

Marginal Distributions



Outline

The Gaussian copula induces dependence via the copula association matrix $\boldsymbol{\Sigma}$.

$$C(\boldsymbol{y}; \boldsymbol{\Sigma}) = \Phi_{\boldsymbol{\Sigma}}[\Phi^{-1}\{F_1(y_1)\}, \ldots, \Phi^{-1}\{F_n(y_n)\}]$$

The Gaussian copula induces dependence via the copula association matrix $\mathbf{\Sigma}$.

$$C(\boldsymbol{y}; \mathbf{\Sigma}) = \Phi_{\mathbf{\Sigma}}[\Phi^{-1}\{F_1(y_1)\}, \ldots, \Phi^{-1}\{F_n(y_n)\}]$$

$$\mathbf{\Sigma}_{ij} = \mathrm{corr}\left(\Phi^{-1}\{F_i(Y_i)\}, \Phi^{-1}\{F_j(Y_j)\}\right)$$

The Gaussian copula induces dependence via the copula association matrix $\boldsymbol{\Sigma}$.

$$C(\boldsymbol{y}; \boldsymbol{\Sigma}) = \Phi_{\boldsymbol{\Sigma}}[\Phi^{-1}\{F_1(y_1)\}, \ldots, \Phi^{-1}\{F_n(y_n)\}]$$

$$\boldsymbol{\Sigma}_{ij} = \text{corr}\left(\Phi^{-1}\{F_i(Y_i)\}, \Phi^{-1}\{F_j(Y_j)\}\right) \neq \text{corr}(Y_i, Y_j)$$

The Gaussian copula induces dependence via the copula association matrix $\mathbf{\Sigma}$.

$$C(\boldsymbol{y}; \mathbf{\Sigma}) = \Phi_{\mathbf{\Sigma}}[\Phi^{-1}\{F_1(y_1)\}, \dots, \Phi^{-1}\{F_n(y_n)\}]$$

$$\mathbf{\Sigma}_{ij} = \text{corr}\left(\Phi^{-1}\{F_i(Y_i)\}, \Phi^{-1}\{F_j(Y_j)\}\right) \neq \text{corr}(Y_i, Y_j)$$

If F_i and F_j are continuous, $\mathbf{\Sigma}_{ij}$ is the **rank** correlation between Y_i and Y_j.

The Gaussian copula induces dependence via the copula association matrix $\boldsymbol{\Sigma}$.

$$C(\boldsymbol{y}; \boldsymbol{\Sigma}) = \Phi_{\boldsymbol{\Sigma}}[\Phi^{-1}\{F_1(y_1)\}, \ldots, \Phi^{-1}\{F_n(y_n)\}]$$

$$\boldsymbol{\Sigma}_{ij} = \text{corr}\left(\Phi^{-1}\{F_i(Y_i)\}, \Phi^{-1}\{F_j(Y_j)\}\right) \neq \text{corr}(Y_i, Y_j)$$

If F_i and F_j are continuous, $\boldsymbol{\Sigma}_{ij}$ is the **rank** correlation between Y_i and Y_j.

Model the elements of $\boldsymbol{\Sigma}$ as a decreasing function of distance.

Variogram Models

Definition: When h_{ij} denotes the distance between locations of observations i and j, the **isotropic variogram** is

$$2\gamma(h_{ij}) = \mathrm{var}(Y_i - Y_j)$$

Variogram Models

Definition: When h_{ij} denotes the distance between locations of observations i and j, the **isotropic variogram** is

$$2\gamma(h_{ij}) = \text{var}(Y_i - Y_j)$$

Exponential Semivariogram:

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ \theta_n + \theta_s[1 - \exp(-h/\theta_r)], & h > 0 \end{cases}$$

Variogram Models

Definition: When h_{ij} denotes the distance between locations of observations i and j, the **isotropic variogram** is

$$2\gamma(h_{ij}) = \text{var}(Y_i - Y_j)$$

Exponential Semivariogram:

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ \theta_n + \theta_s[1 - \exp(-h/\theta_r)], & h > 0 \end{cases}$$

θ_n is the **nugget**

Variogram Models

Definition: When h_{ij} denotes the distance between locations of observations i and j, the **isotropic variogram** is

$$2\gamma(h_{ij}) = \text{var}(Y_i - Y_j)$$

Exponential Semivariogram:

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ \theta_n + \theta_s[1 - \exp(-h/\theta_r)], & h > 0 \end{cases}$$

θ_n is the **nugget**

θ_s is the **partial sill**

Variogram Models

Definition: When h_{ij} denotes the distance between locations of observations i and j, the **isotropic variogram** is

$$2\gamma(h_{ij}) = \text{var}(Y_i - Y_j)$$

Exponential Semivariogram:

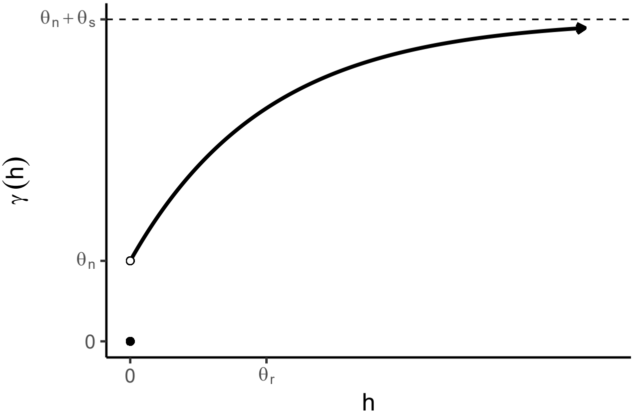
$$\gamma(h) = \begin{cases} 0, & h = 0 \\ \theta_n + \theta_s[1 - \exp(-h/\theta_r)], & h > 0 \end{cases}$$

θ_n is the **nugget**

θ_s is the **partial sill**

θ_r is the **range**

Exponential Semivariogram



Covariance

If $\text{var}(Y_i) = \text{var}(Y_j)$, then $2\gamma(h_{ij}) = 2\,\text{var}(Y_i) - 2\,\text{cov}(Y_i, Y_j)$

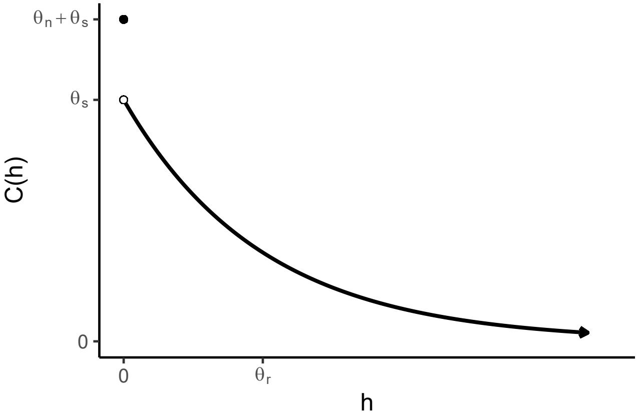
Covariance

If $\text{var}(Y_i) = \text{var}(Y_j)$, then $2\gamma(h_{ij}) = 2\,\text{var}(Y_i) - 2\,\text{cov}(Y_i, Y_j)$

Exponential covariance function:

$$C(h) = \begin{cases} \theta_n + \theta_s, & h = 0 \\ \theta_s \exp(-h/\theta_r), & h > 0 \end{cases}$$

Exponential Covariance Function

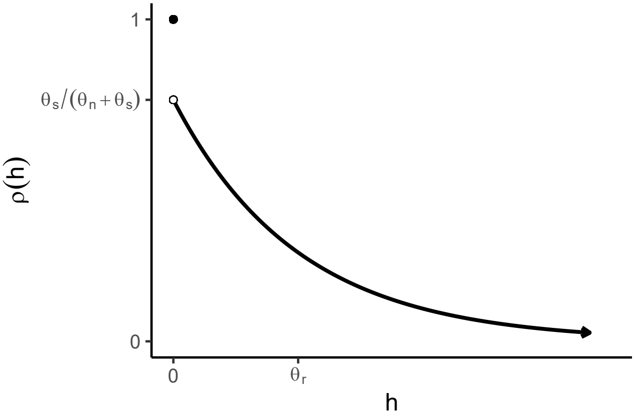


Correlation

Exponential correlation function:

$$\rho(h) = \begin{cases} 1, & h = 0 \\ \frac{\theta_s}{\theta_s + \theta_n} \exp(-h/\theta_r), & h > 0 \end{cases}$$

Exponential Correlation Function



Spatial Gaussian Copula Model

Let $\boldsymbol{Y} = [Y_1, \ldots Y_n]$ have cdf

$$F(\boldsymbol{y}; \boldsymbol{\Sigma}) = \Phi_{\boldsymbol{\Sigma}}[\Phi^{-1}\{F_1(y_1)\}, \ldots, \Phi^{-1}\{F_n(y_n)\}].$$

Spatial Gaussian Copula Model

Let $\boldsymbol{Y} = [Y_1, \ldots Y_n]$ have cdf

$$F(\boldsymbol{y}; \boldsymbol{\Sigma}) = \Phi_{\boldsymbol{\Sigma}}[\Phi^{-1}\{F_1(y_1)\}, \ldots, \Phi^{-1}\{F_n(y_n)\}].$$

Association matrix $\boldsymbol{\Sigma}$ has ijth element

$$\Sigma_{ij} = \rho(h_{ij}) = \begin{cases} 1, & h_{ij} = 0 \\ \alpha_N \exp(-h_{ij}/\alpha_R), & h_{ij} > 0 \end{cases},$$

where h_{ij} denotes the Euclidean distance between locations of observations i and j.

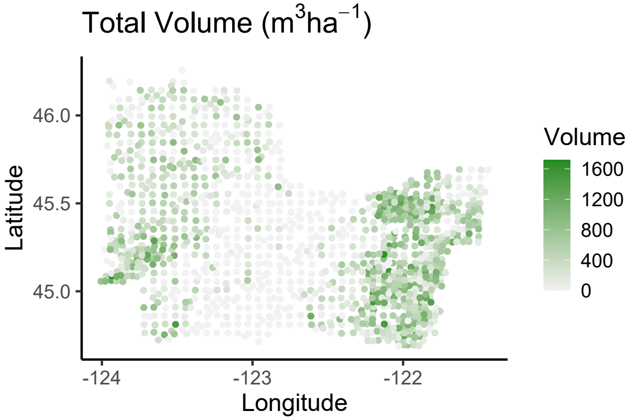
Zero Inflated Continuous Marginals

$Y_i \sim F_i$ with

$$F_i(y) = \begin{cases} 0, & y < 0 \\ y \cdot p_i/\epsilon, & 0 \leq y < \epsilon \\ p_i + (1 - p_i)F_{\text{lnorm}}(y - \epsilon; \mu_i, \sigma^2), & y \geq \epsilon \end{cases}$$

where p_i and μ_i may depend on covariates.

Survey Unit 1 Volume Map



$n = 1224$ plots

298 with zero total volume

Logistic Submodel

Logistic regression model for the Bernoulli process:

$$
\begin{aligned}
B_i &= \begin{cases} 1, & \text{plot } i \text{ total volume} = 0 \\ 0, & \text{otherwise} \end{cases} \\
B_i &\sim \text{Bernoulli}(p_i) \\
\text{logit}(p_i) &= \boldsymbol{X}_i \boldsymbol{\beta} \\
\boldsymbol{X}_i &= \text{row vector of covariates}
\end{aligned}
$$

Lognormal Submodel

Log-linear regression model for the 926 non-0 volume observations:

$$
\begin{aligned}
Y_i &= \text{total volume in } i\text{th plot, if positive} \\
\log(\boldsymbol{Y}) &\sim N(\boldsymbol{\mu}, \sigma^2) \\
\boldsymbol{\mu} &= \boldsymbol{X}_y \gamma
\end{aligned}
$$

where $\boldsymbol{X}_y$ is a design matrix of covariates.

Potential covariates:

`forind`	indicator of forest
`annpre`	mean annual precipitation
`anntmp`	mean annual temperature
`smrtp`	moisture stress during growing season
`ndvi`	vegetation greenness
`tc1`	brightness
`tc2`	greenness
`tc3`	wetness

Logistic Model Fit

From R's `glm` function.

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.7570     0.2355 -11.709  < 2e-16 ***
forind        -0.5957     0.1425  -4.181 2.90e-05 ***
tc1            1.2641     0.2315   5.460 4.77e-08 ***
tc2           -0.6842     0.2146  -3.188 0.001435 **
tc3           -0.9236     0.1861  -4.963 6.95e-07 ***
annpre        -1.2260     0.3363  -3.646 0.000267 ***
anntmp         2.1818     0.3405   6.407 1.49e-10 ***
smrtp         -1.3573     0.4446  -3.053 0.002268 **
ndvi          -0.7923     0.2051  -3.862 0.000112 ***
forind:tc2    -0.2492     0.1154  -2.160 0.030756 *
forind:tc3    -0.4386     0.1351  -3.247 0.001167 **
```

Lognormal Model Fit

From R's `lm` function.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.14220    0.05453  94.308  < 2e-16 ***
tc1          -0.57893    0.10593  -5.465 5.96e-08 ***
tc2          -0.13481    0.07386  -1.825  0.06828 .
tc3           0.82963    0.08732   9.501  < 2e-16 ***
annpre        0.17192    0.04515   3.808  0.00015 ***
anntmp        0.11513    0.04663   2.469  0.01373 *
tc1:ndvi      0.08206    0.06750   1.216  0.22437
tc2:ndvi      0.09974    0.05807   1.718  0.08617 .
anntmp:ndvi -0.31491    0.05041  -6.247 6.38e-10 ***
tc3:smrtp     0.27859    0.08820   3.159  0.00164 **
tc3:annpre    0.02789    0.08764   0.318  0.75040
```

Fitting the Spatial Model

Copula association matrix $\boldsymbol{\Sigma}$ is the spatial correlation matrix of $\Phi^{-1}\{F_1(Y_1)\}, \ldots, \Phi^{-1}\{F_n(Y_n)\}$.

Fitting the Spatial Model

Copula association matrix $\boldsymbol{\Sigma}$ is the spatial correlation matrix of $\Phi^{-1}\{F_1(Y_1)\}, \ldots, \Phi^{-1}\{F_n(Y_n)\}$.

Estimate variogram of $\Phi^{-1}\{\widehat{F}_1(Y_1)\}, \ldots, \Phi^{-1}\{\widehat{F}_n(Y_n)\}$, where

$$\widehat{F}_i(y) = \begin{cases} 0, & y < 0 \\ y \cdot \widehat{p}_i/\epsilon, & 0 \leq y < \epsilon \\ \widehat{p}_i + (1 - \widehat{p}_i)F_{\text{lnorm}}(y - \epsilon; \widehat{\mu}_i, \widehat{\sigma}^2), & y \geq \epsilon \end{cases}$$

Fitting the Spatial Model

Copula association matrix $\boldsymbol{\Sigma}$ is the spatial correlation matrix of $\Phi^{-1}\{F_1(Y_1)\}, \ldots, \Phi^{-1}\{F_n(Y_n)\}$.

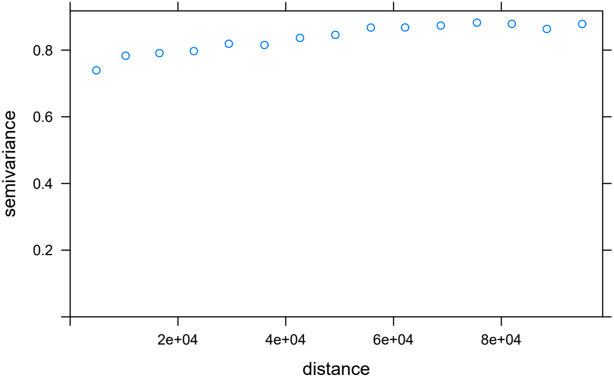
Estimate variogram of $\Phi^{-1}\{\widehat{F}_1(Y_1)\}, \ldots, \Phi^{-1}\{\widehat{F}_n(Y_n)\}$, where

$$\widehat{F}_i(y) = \begin{cases} 0, & y < 0 \\ y \cdot \widehat{p}_i/\epsilon, & 0 \leq y < \epsilon \\ \widehat{p}_i + (1 - \widehat{p}_i) F_{\text{lnorm}}(y - \epsilon; \widehat{\mu}_i, \widehat{\sigma}^2), & y \geq \epsilon \end{cases}$$

Choose $\epsilon < \min\{Y_i | Y_i > 0\}$

Empirical Variogram

Calculate empirical variogram of $\Phi^{-1}\{\widehat{F_1}(y_1)\}, \ldots, \Phi^{-1}\{\widehat{F_n}(y_n)\}$ using R `gstat` package.

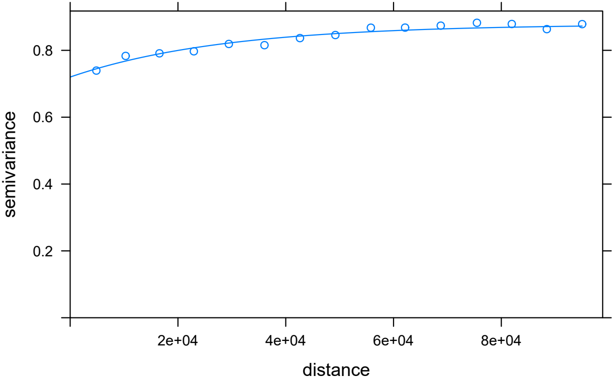


Fitted Exponential Variogram

	model	psill	range
1	Nug	0.7201172	0.00
2	Exp	0.1581205	28559.81

Fitted Exponential Variogram

```
  model     psill    range
1   Nug 0.7201172     0.00
2   Exp 0.1581205 28559.81
```



Estimated Exponential Correlation Function

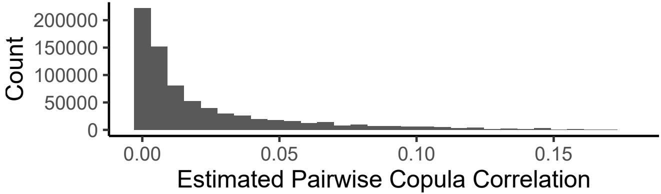
$$\widehat{\alpha_N} = \frac{0.1581205}{0.1581205 + 0.7201172} \approx 0.18$$
$$\widehat{\alpha_R} \approx 28560$$

For plot i and plot j separated by h_{ij} meters,
$\widehat{\text{corr}}\left[\Phi^{-1}\{F_i(y_i)\}, \Phi^{-1}\{F_j(y_j)\}\right] = 0.18\exp(-h_{ij}/28560)$.

Estimated Exponential Correlation Function

$$\widehat{\alpha_N} = \frac{0.1581205}{0.1581205 + 0.7201172} \approx 0.18$$
$$\widehat{\alpha_R} \approx 28560$$

For plot i and plot j separated by h_{ij} meters,
$\widehat{\text{corr}}\left[\Phi^{-1}\{F_i(y_i)\}, \Phi^{-1}\{F_j(y_j)\}\right] = 0.18 \exp(-h_{ij}/28560)$.

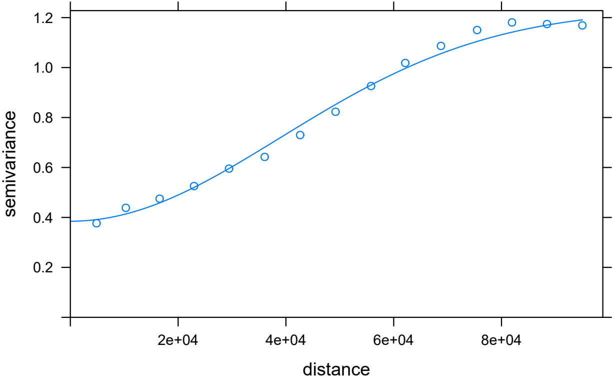


Fitted Variogram for Intercept-only Models

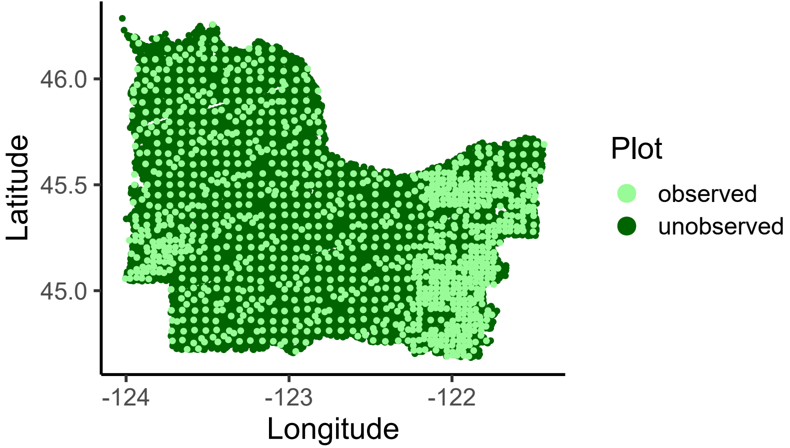
$$\begin{aligned} \text{logit}(p_i) &= \beta_0 \\ \mu_i &= \mu \end{aligned}$$

Fitted Variogram for Intercept-only Models

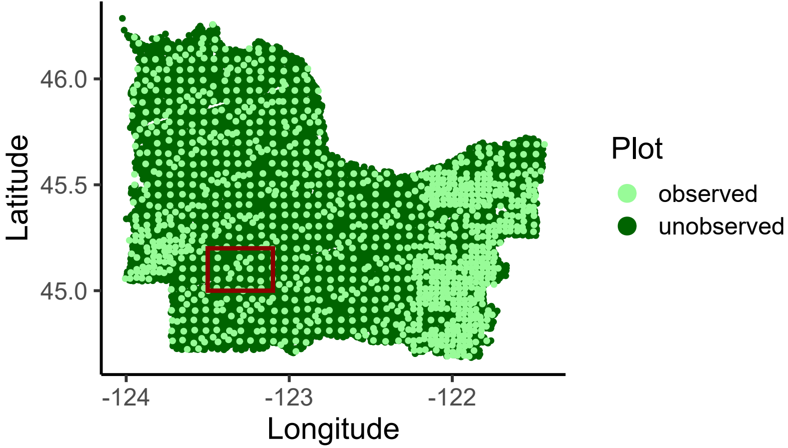
$$\text{logit}(p_i) = \beta_0$$
$$\mu_i = \mu$$



Survey Unit 0



Block Kriging



Kriging With Normal Data

Suppose

$$\begin{bmatrix} \boldsymbol{Z}_{\text{o}} \\ \boldsymbol{Z}_{\text{u}} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_{\text{o}} \\ \boldsymbol{\mu}_{\text{u}} \end{bmatrix}, \begin{bmatrix} \Sigma_{\text{o}} & \Sigma_{\text{ou}} \\ \Sigma_{\text{ou}}' & \Sigma_{\text{u}} \end{bmatrix} \right)$$

where

$$\begin{aligned} \boldsymbol{Z}_{\text{o}} &\sim N(\boldsymbol{\mu}_{\text{o}}, \Sigma_{\text{o}}) \text{ are the observed data} \\ \boldsymbol{Z}_{\text{u}} &\sim N(\boldsymbol{\mu}_{\text{u}}, \Sigma_{\text{u}}) \text{ are the unobserved data} \\ \Sigma_{\text{ou}} &= \text{cov}(\boldsymbol{Z}_{\text{o}}, \boldsymbol{Z}_{\text{u}}) \text{ is the cross-covariance matrix} \end{aligned}$$

Predicting $\boldsymbol{Z}_{\mathrm{u}}$

Then

$$
\begin{aligned}
E(\boldsymbol{Z}_{\mathrm{u}}|\boldsymbol{Z}_{\mathrm{o}}) &= \boldsymbol{\mu}_{\mathrm{u}} + \Sigma'_{\mathrm{ou}}\Sigma_{\mathrm{o}}^{-1}(\boldsymbol{Z}_{\mathrm{o}} - \boldsymbol{\mu}_{\mathrm{o}}) \\
\mathrm{var}(\boldsymbol{Z}_{\mathrm{u}}|\boldsymbol{Z}_{\mathrm{o}}) &= \Sigma_{\mathrm{u}} - \Sigma'_{\mathrm{ou}}\Sigma_{\mathrm{o}}^{-1}\Sigma_{\mathrm{ou}}
\end{aligned}
$$

Predicting $\boldsymbol{Z}_{\text{u}}$

Then

$$
\begin{aligned}
E(\boldsymbol{Z}_{\text{u}}|\boldsymbol{Z}_{\text{o}}) &= \boldsymbol{\mu}_{\text{u}} + \Sigma_{\text{ou}}^{\prime}\Sigma_{\text{o}}^{-1}(\boldsymbol{Z}_{\text{o}} - \boldsymbol{\mu}_{\text{o}}) \\
\text{var}(\boldsymbol{Z}_{\text{u}}|\boldsymbol{Z}_{\text{o}}) &= \Sigma_{\text{u}} - \Sigma_{\text{ou}}^{\prime}\Sigma_{\text{o}}^{-1}\Sigma_{\text{ou}}
\end{aligned}
$$

Empirical Best Linear Unbiased Predictor is

$$
\widehat{\boldsymbol{Z}}_{\text{u}} = \widehat{\boldsymbol{\mu}}_{\text{u}} + \widehat{\Sigma}_{\text{ou}}^{\prime}\widehat{\Sigma}_{\text{o}}^{-1}(\boldsymbol{Z}_{\text{o}} - \widehat{\boldsymbol{\mu}}_{\text{o}}),
$$

where $\widehat{\boldsymbol{\mu}}_{\text{u}}$ and $\widehat{\boldsymbol{\mu}}_{\text{o}}$ are weighted least squares estimates.

Predicting $\boldsymbol{Z}_{\text{U}}$

Then

$$E(\boldsymbol{Z}_{\text{U}}|\boldsymbol{Z}_{\text{O}}) = \boldsymbol{\mu}_{\text{U}} + \Sigma'_{\text{OU}}\Sigma_{\text{O}}^{-1}(\boldsymbol{Z}_{\text{O}} - \boldsymbol{\mu}_{\text{O}})$$

$$\text{var}(\boldsymbol{Z}_{\text{U}}|\boldsymbol{Z}_{\text{O}}) = \Sigma_{\text{U}} - \Sigma'_{\text{OU}}\Sigma_{\text{O}}^{-1}\Sigma_{\text{OU}}$$

Empirical Best Linear Unbiased Predictor is

$$\widehat{\boldsymbol{Z}}_{\text{U}} = \widehat{\boldsymbol{\mu}}_{\text{U}} + \widehat{\Sigma}'_{\text{OU}}\widehat{\Sigma}_{\text{O}}^{-1}(\boldsymbol{Z}_{\text{O}} - \widehat{\boldsymbol{\mu}}_{\text{O}}),$$

where $\widehat{\boldsymbol{\mu}}_{\text{U}}$ and $\widehat{\boldsymbol{\mu}}_{\text{O}}$ are weighted least squares estimates.

Block kriging estimate of total is $\widehat{T} = \mathbf{1}'\widehat{\boldsymbol{Z}}_{\text{U}} + \mathbf{1}'\boldsymbol{Z}_{\text{O}}$.

Predicting $\boldsymbol{Z}_{\text{U}}$

Then

$$
\begin{aligned}
E(\boldsymbol{Z}_{\text{U}}|\boldsymbol{Z}_{\text{O}}) &= \boldsymbol{\mu}_{\text{U}} + \Sigma_{\text{OU}}'\Sigma_{\text{O}}^{-1}(\boldsymbol{Z}_{\text{O}} - \boldsymbol{\mu}_{\text{O}}) \\
\text{var}(\boldsymbol{Z}_{\text{U}}|\boldsymbol{Z}_{\text{O}}) &= \Sigma_{\text{U}} - \Sigma_{\text{OU}}'\Sigma_{\text{O}}^{-1}\Sigma_{\text{OU}}
\end{aligned}
$$

Empirical Best Linear Unbiased Predictor is

$$
\widehat{\boldsymbol{Z}}_{\text{U}} = \widehat{\boldsymbol{\mu}}_{\text{U}} + \widehat{\Sigma}_{\text{OU}}'\widehat{\Sigma}_{\text{O}}^{-1}(\boldsymbol{Z}_{\text{O}} - \widehat{\boldsymbol{\mu}}_{\text{O}}),
$$

where $\widehat{\boldsymbol{\mu}}_{\text{U}}$ and $\widehat{\boldsymbol{\mu}}_{\text{O}}$ are weighted least squares estimates.

Block kriging estimate of total is $\widehat{T} = \mathbf{1}'\widehat{\boldsymbol{Z}}_{\text{U}} + \mathbf{1}'\boldsymbol{Z}_{\text{O}}$.

$\text{var}(\widehat{T} - T)$ depends on Σ.

Block Kriging with the Y's

Let $\boldsymbol{Y}_{\text{O}} = [Y_1, \ldots, Y_n]$ be the vector of responses on the observed plots and $\boldsymbol{Y}_{\text{U}} = [Y_{n+1}, \ldots, Y_{n+m}]$ be the vector of responses on the unobserved plots.

Definitions:

$$\begin{aligned} Z_i &= \Phi^{-1}\{F_i(Y_i)\}, i = 1, \ldots, n+m \\ \boldsymbol{Z}_{\text{O}} &= Z_1, \ldots, Z_n \\ \boldsymbol{Z}_{\text{U}} &= Z_{n+1}, \ldots, Z_{n+m} \end{aligned}$$

Block Kriging with the Y's

Let $\boldsymbol{Y}_{\text{O}} = [Y_1, \ldots, Y_n]$ be the vector of responses on the observed plots and $\boldsymbol{Y}_{\text{U}} = [Y_{n+1}, \ldots, Y_{n+m}]$ be the vector of responses on the unobserved plots.

Definitions:

$$
\begin{aligned}
Z_i &= \Phi^{-1}\{F_i(Y_i)\}, i = 1, \ldots, n+m \\
\boldsymbol{Z}_{\text{O}} &= Z_1, \ldots, Z_n \\
\boldsymbol{Z}_{\text{U}} &= Z_{n+1}, \ldots, Z_{n+m} \\
\widehat{Z}_i &= \Phi^{-1}\{\widehat{F}_i(Y_i)\}, i = 1, \ldots, n \\
\widehat{\boldsymbol{Z}}_{\text{O}} &= \widehat{Z}_1, \ldots, \widehat{Z}_n
\end{aligned}
$$

Exploit the Copula Normalizing Transformation

Copula model:

$$\begin{bmatrix} \boldsymbol{Z}_{\mathrm{o}} \\ \boldsymbol{Z}_{\mathrm{u}} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0}_n \\ \mathbf{0}_m \end{bmatrix}, \begin{bmatrix} \Sigma_{\mathrm{o}} & \Sigma_{\mathrm{ou}} \\ \Sigma_{\mathrm{ou}}' & \Sigma_{\mathrm{u}} \end{bmatrix} \right)$$

Exploit the Copula Normalizing Transformation

Copula model:

$$\begin{bmatrix} \boldsymbol{Z}_{\mathrm{o}} \\ \boldsymbol{Z}_{\mathrm{u}} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0}_n \\ \mathbf{0}_m \end{bmatrix}, \begin{bmatrix} \Sigma_{\mathrm{o}} & \Sigma_{\mathrm{ou}} \\ \Sigma'_{\mathrm{ou}} & \Sigma_{\mathrm{u}} \end{bmatrix} \right)$$

Predict $\boldsymbol{Z}_{\mathrm{u}}$ as

$$\widehat{\boldsymbol{Z}}_{\mathrm{u}} = \widehat{\Sigma}'_{\mathrm{ou}} \widehat{\Sigma}_{\mathrm{o}}^{-1} \widehat{\boldsymbol{Z}}_{\mathrm{o}}.$$

Exploit the Copula Normalizing Transformation

Copula model:

$$\begin{bmatrix} \boldsymbol{Z}_{\text{o}} \\ \boldsymbol{Z}_{\text{u}} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0}_n \\ \boldsymbol{0}_m \end{bmatrix}, \begin{bmatrix} \Sigma_{\text{o}} & \Sigma_{\text{ou}} \\ \Sigma_{\text{ou}}' & \Sigma_{\text{u}} \end{bmatrix} \right)$$

Predict $\boldsymbol{Z}_{\text{u}}$ as

$$\widehat{\boldsymbol{Z}}_{\text{u}} = \widehat{\Sigma}_{\text{ou}}' \widehat{\Sigma}_{\text{o}}^{-1} \widehat{\boldsymbol{Z}}_{\text{o}}.$$

Then let $\widehat{Y}_i = \widehat{F}_i^{-1}\{\Phi(\widehat{Z}_i)\}, i = 1, \ldots, n + m$

Exploit the Copula Normalizing Transformation

Copula model:

$$\begin{bmatrix} \boldsymbol{Z}_{\text{o}} \\ \boldsymbol{Z}_{\text{u}} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0}_n \\ \boldsymbol{0}_m \end{bmatrix}, \begin{bmatrix} \Sigma_{\text{o}} & \Sigma_{\text{ou}} \\ \Sigma'_{\text{ou}} & \Sigma_{\text{u}} \end{bmatrix} \right)$$

Predict $\boldsymbol{Z}_{\text{u}}$ as

$$\widehat{\boldsymbol{Z}}_{\text{u}} = \widehat{\Sigma}'_{\text{ou}} \widehat{\Sigma}_{\text{o}}^{-1} \widehat{\boldsymbol{Z}}_{\text{o}}.$$

Then let $\widehat{Y}_i = \widehat{F}_i^{-1}\{\Phi(\widehat{Z}_i)\}, i = 1, \ldots, n + m$

Estimate total T as $\sum_{i=1}^{n+m} \widehat{Y}_i$.

Bad News

This doesn’t work.

Bad News

This doesn't work.

$\widehat{\boldsymbol{Z}}_{\mathrm{u}}$ estimates $E(\boldsymbol{Z}_{\mathrm{u}}|\boldsymbol{Z}_{\mathrm{o}})$, and $F_i^{-1}\{\Phi(E(Z_i|\boldsymbol{Z}_{\mathrm{o}}))\} \neq E(Y_i|\boldsymbol{Y}_{\mathrm{o}})$.

Bad News

This doesn't work.

$\widehat{\boldsymbol{Z}}_{\text{U}}$ estimates $E(\boldsymbol{Z}_{\text{U}}|\boldsymbol{Z}_{\text{O}})$, and $F_i^{-1}\{\Phi(E(Z_i|\boldsymbol{Z}_{\text{O}}))\} \neq E(Y_i|\boldsymbol{Y}_{\text{O}})$.

Prediction does not account for variance of parameter estimates.

Proposed Solution–Parametric Bootstrap

- Estimate parameters $\boldsymbol{\theta} = [\alpha_N, \alpha_R, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma^2]$ by maximizing copula likelihood.

Proposed Solution–Parametric Bootstrap

- Estimate parameters $\boldsymbol{\theta} = [\alpha_N, \alpha_R, \boldsymbol{\theta}, \boldsymbol{\gamma}, \sigma^2]$ by maximizing copula likelihood.
- Generate $k = 1, \ldots, N_b$ realizations of $\widehat{\boldsymbol{\theta}}$ from estimated asymptotic distribution.

Proposed Solution–Parametric Bootstrap

- Estimate parameters $\boldsymbol{\theta} = [\alpha_N, \alpha_R, \boldsymbol{\theta}, \gamma, \sigma^2]$ by maximizing copula likelihood.
- Generate $k = 1, \ldots, N_b$ realizations of $\widehat{\boldsymbol{\theta}}$ from estimated asymptotic distribution.
- For each realization of $\widehat{\boldsymbol{\theta}}$,
  - Generate $\widehat{\boldsymbol{Z}}_{\mathrm{u}} \sim N(\mathbf{0}, \widehat{\Sigma}_{\mathrm{u}} - \widehat{\Sigma}'_{\mathrm{ou}}\widehat{\Sigma}_{\mathrm{o}}^{-1}\widehat{\Sigma}_{\mathrm{ou}})$.

Proposed Solution–Parametric Bootstrap

- Estimate parameters $\boldsymbol{\theta} = [\alpha_N, \alpha_R, \boldsymbol{\theta}, \gamma, \sigma^2]$ by maximizing copula likelihood.
- Generate $k = 1, \ldots, N_b$ realizations of $\widehat{\boldsymbol{\theta}}$ from estimated asymptotic distribution.
- For each realization of $\widehat{\boldsymbol{\theta}}$,
  - Generate $\widehat{\boldsymbol{Z}}_{\mathrm{u}} \sim N(\mathbf{0}, \widehat{\Sigma}_{\mathrm{u}} - \widehat{\Sigma}'_{\mathrm{ou}}\widehat{\Sigma}_{\mathrm{o}}^{-1}\widehat{\Sigma}_{\mathrm{ou}})$.
  - Transform elements of $\widehat{\boldsymbol{Z}}_{\mathrm{u}}$ to data scale: $\widehat{Y}_i = \widehat{F}_i^{-1}\{\Phi(\widehat{Z}_i)\}$, where $\widehat{F}_i$ is based on the realization of $\widehat{\boldsymbol{\theta}}$.

Proposed Solution–Parametric Bootstrap

- Estimate parameters $\boldsymbol{\theta} = [\alpha_N, \alpha_R, \boldsymbol{\theta}, \gamma, \sigma^2]$ by maximizing copula likelihood.
- Generate $k = 1, \ldots, N_b$ realizations of $\widehat{\boldsymbol{\theta}}$ from estimated asymptotic distribution.
- For each realization of $\widehat{\boldsymbol{\theta}}$,
  - Generate $\widehat{\boldsymbol{Z}}_{\text{U}} \sim N(\mathbf{0}, \widehat{\Sigma}_{\text{U}} - \widehat{\Sigma}'_{\text{OU}} \widehat{\Sigma}_{\text{O}}^{-1} \widehat{\Sigma}_{\text{OU}})$.
  - Transform elements of $\widehat{\boldsymbol{Z}}_{\text{U}}$ to data scale: $\widehat{Y}_i = \widehat{F}_i^{-1}\{\Phi(\widehat{Z}_i)\}$, where $\widehat{F}_i$ is based on the realization of $\widehat{\boldsymbol{\theta}}$.
  - Calculate $\widehat{T} = \mathbf{1}'\widehat{\boldsymbol{Y}}_{\text{U}} + \mathbf{1}'\boldsymbol{Y}_{\text{O}}$

Proposed Solution–Parametric Bootstrap

- Estimate parameters $\boldsymbol{\theta} = [\alpha_N, \alpha_R, \boldsymbol{\theta}, \gamma, \sigma^2]$ by maximizing copula likelihood.
- Generate $k = 1, \ldots, N_b$ realizations of $\widehat{\boldsymbol{\theta}}$ from estimated asymptotic distribution.
- For each realization of $\widehat{\boldsymbol{\theta}}$,
  - Generate $\widehat{\boldsymbol{Z}}_{\text{u}} \sim N(\mathbf{0}, \widehat{\Sigma}_{\text{u}} - \widehat{\Sigma}'_{\text{ou}}\widehat{\Sigma}_{\text{o}}^{-1}\widehat{\Sigma}_{\text{ou}})$.
  - Transform elements of $\widehat{\boldsymbol{Z}}_{\text{u}}$ to data scale: $\widehat{Y}_i = \widehat{F}_i^{-1}\{\Phi(\widehat{Z}_i)\}$, where $\widehat{F}_i$ is based on the realization of $\widehat{\boldsymbol{\theta}}$.
  - Calculate $\widehat{T} = \mathbf{1}'\widehat{\boldsymbol{Y}}_{\text{u}} + \mathbf{1}'\boldsymbol{Y}_{\text{o}}$
- This yields a bootstrapped distribution of $\widehat{T}$'s. Take the median as the point prediction of the block total, and take $\alpha/2$ and $1 - \alpha/2$ quantiles as lower and upper $1 - \alpha$ prediction limits.

Preliminary Simulations

Covariate vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ iid Uniform$(0, 1)$.

Bernoulli model: $\text{logit}(p_i) = 1 - 3X_{1\,i}$

Lognormal model: $\log(Y_i) \sim N(1 + 5X_{2\,i}, 1)$

Preliminary Simulations

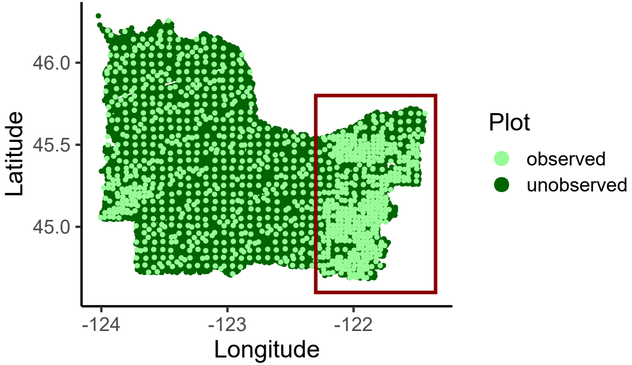
Covariate vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ iid Uniform$(0, 1)$.

Bernoulli model: $\text{logit}(p_i) = 1 - 3X_{1\,i}$

Lognormal model: $\log(Y_i) \sim N(1 + 5X_{2\,i}, 1)$

This gives approximately 40% 0's.

Locations



481 observed plots
3047 unobserved plots

Exponential Spatial Dependence

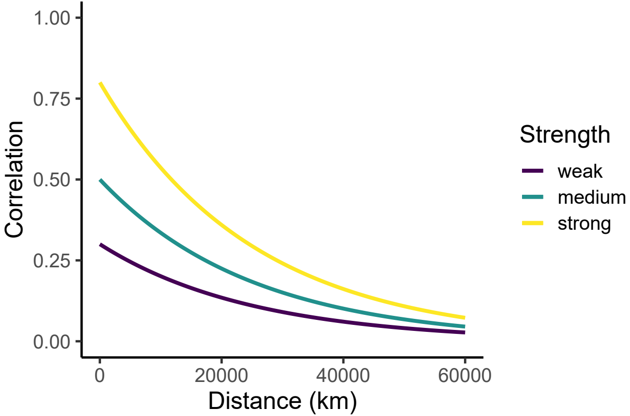
Range: $\alpha_R = 25000$

Nugget: $\alpha_N = 0.3, 0.5, 0.8$ (weak, medium, strong dependence)

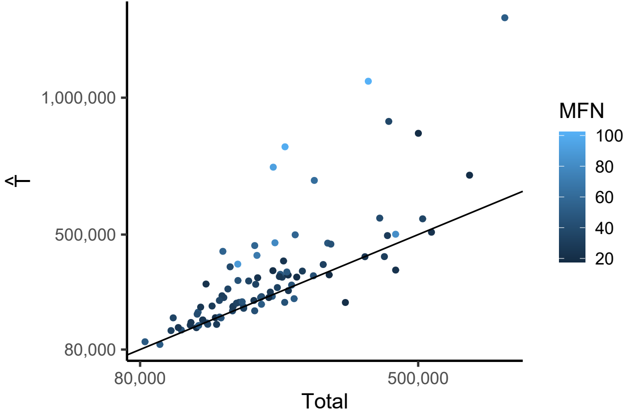
Exponential Spatial Dependence

Range: $\alpha_R = 25000$

Nugget: $\alpha_N = 0.3, 0.5, 0.8$ (weak, medium, strong dependence)

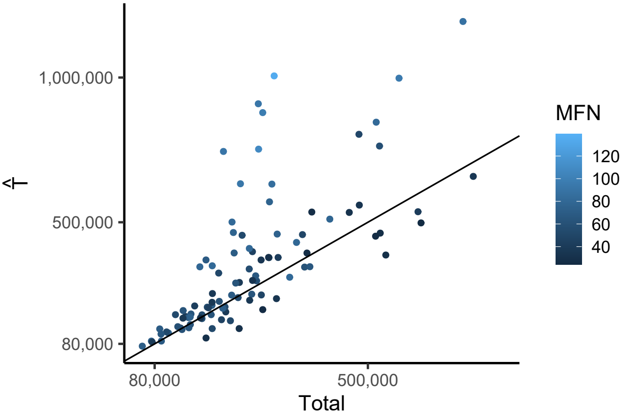


Results for $\alpha_N = 0.3$

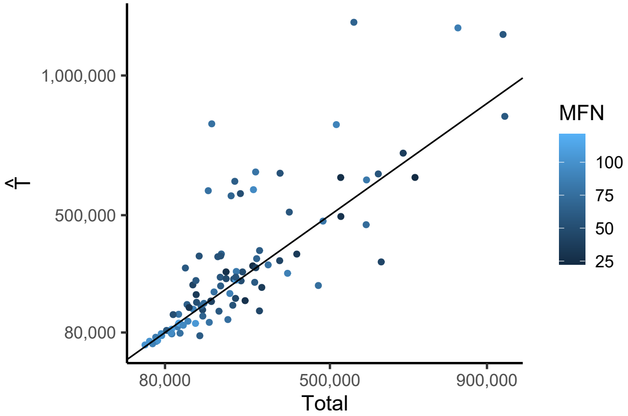


MFN is the median Frobenius norm of $\Sigma - \widehat{\Sigma}$ over the bootstrapped sample.

Results for $\alpha_N = 0.5$



Results for $\alpha_N = 0.8$



Summary

- Gaussian copula models spatial dependence.
- Continuous zero-inflated lognormal marginal models accommodate large percentage of zeros.
- For FIA data, marginal models account for most of the spatial pattern, so spatial prediction not needed.

Next Steps

- Simulations
- Model assumptions
- Big data

Thanks

Funding for this research was provided by USDA Forest Service Agreement 17-MU-11261959-061.

Many thanks to Lisa Wilson.