Comparison of Gaussian copula and random forests

in zero-inflated spatial prediction

Nick Sun

Oregon State University, Department of Statistics

June 4, 2020

Outline

Forestry Inventory

- Forestry inventory is a critical part of monitoring and servicing ecosystems and often involves statistical estimation of quantities such as total wood volume
- Since forests can cover enormous areas over rough terrain, it is often not possible to sample certain areas of forests due to physical, budgetary, or time constraints
- However, forestry data is often zero-inflated, heavily skewed, and spatially dependent, making it difficult to model using traditional statistical and geostatistical models

Original Forestry Data

- The forestry inventory data used here was made available by the USDA Forestry Inventory and Analysis program, containing 13 variables of interest across 1224 plots of land in northwest Oregon.

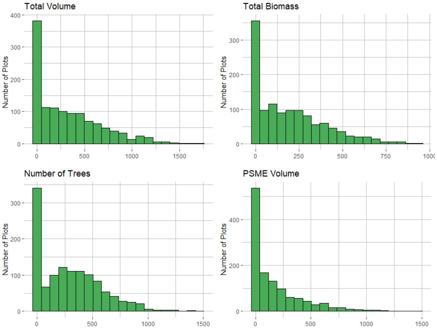

Figure: Histograms of forestry inventory variables.

Simulated Data

We simulated $m = 1000$ datasets of size $n = 1224$ for total timber volume and hemlock volume, two common variables of interest in forestry inventory applications. Previous work used a zero-inflated Gamma distribution to model these variables[4].

- We create simulated datasets by generating multivariate normal observations with the sample correlation matrix from the original data
- We then backtransform using the quantile function of the zero-inflated gamma function that was found to fit the original data

Resampled Data

- We also generate training datasets by sampling rows without replacement from the orignal data with the remaining rows serving as a test set
- For models trained on these resampled datasets, we will be able to use covariates present in the Oregon dataset in our models, such as average annual precipitation

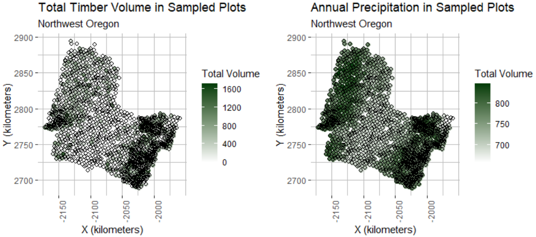

Figure: Alber's Equal Area Conic projection used here.

New Spatial Models

- Two new techniques have been proposed to estimate spatially dependent data: spatial Gaussian copula and spatial random forests
- These simulations will compare the predictive performance of these new models and traditional kriging in different spatial prediction scenarios

Kriging

- In geostatistics, kriging is a method of spatial interpolation where values at unobserved locations are estimated using a weighted sum of known values
- In particular, if the data is normally distributed and satisfies *second order stationarity*, *i.e.* the covariances of points is a function only of the distance between the points and not the specific physical location of the points themselves, then kriging is the *best linear unbiased estimator*[2]

$$\hat{y}_K(s_0) = w(s_0)^T y$$

Spatial Gaussian Copula

- Copulas are multivariate cumulative distribution functions where each variable has a standard uniform marginal distribution
- An important copula result:

Theorem (Sklar)

Any n-dimensional multivarite cumulative distribution function $G(\vec{X})$ of a random vector $\vec{X} = (X_1, \ldots, X_n)$ can be expressed in terms of the marginal cumulative distribution functions $F_i(X_i)$ and a copula function $C : [0,1]^n \to [0,1]$ such that

$$G(\vec{X}) = C(F_1(X_1), \ldots, F_n(X_n))$$

New Spatial Models

- Madsen[4] proposed a spatial Gaussian copula

$$G(\vec{V}, \Sigma) = \Phi_{\Sigma}(\Phi^{-1}(F_1(v_1)), \ldots, \Phi^{-1}(F_n(v_n)))$$

  where the correlation matrix Σ is chosen such that it represents the spatial relationships between each of the data points.
- Differentiating the above copula yield the joint density function of the spatially dependent data

$$g(\vec{V}) = \|\Sigma\|^{1/2} \exp\left(-\frac{1}{2} z^T (\Sigma^{-1} - I_n) z\right) \prod_{i=1}^{m} f_i(y_i)$$

  where $z = (\Phi^{-1}(F_1(y_1)), \ldots, \Phi^{-1}(F_n(y_n)))$.
- This copula will be able to incorporate the spatial dependency structure with the appropriate selection of F and Σ.

Spatial Gaussian Copula

A common choice for spatial correlation matrix Σ has i,jth entry equal to the value of the exponential correlogram function

$$\Sigma_{ij}(\theta) = \begin{cases} \theta_0 \exp(-h_{ij}\theta_1) & \text{for } i \neq j \\ 1 & \text{when } i = j \end{cases}$$

where h_{ij} is the distance between the locations y_i and y_j, $0 < \theta_0 \leq 1$ is the nugget parameter describing the variation of the data at $h = 0$, and $\theta_1 > 0$ is the decay parameter.

Marginal Distributions for Copula Model

An appropriate F function would be one which can handle semicontinuous data. Previous work used a zero-inflated gamma function on cube-root transformed response data.

$$f(x) = \begin{cases} 0 & \text{w.p } p \\ \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right) & \text{w.p. } 1-p \end{cases}$$

where $p \sim \text{Bernoulli}(\pi)$

- The cube root transformation was used to make the continuous component less skewed
- Zero values were transformed to uniform random variables sampled from a $U(0, \epsilon)$ distribution where ϵ is the smallest nonzero value in the observed dataset

Spatial Random Forests

- The generic random forest is popular a machine learning algorithm which creates an ensemble of decision trees from the original data using *bootstrap aggregation* and *feature bagging*
- One of the notable advantages of using a machine learning algorithm like random forests is that no statistical assumptions are required
- Random forests have been used in spatial prediction, but the spatial information is often disregarded[1].

Spatial Random Forests

- The **RFsp** `R` package (built on top of **ranger**) introduces the spatial random forest:

$$Y(s) = f(X_G, X_R, X_P)$$

  where X_G are covariates based on geographic proximity, and X_R are surface reflectance covariates, and X_P and process-based covariates
- Essentially, spatial dependence is modeled by training the random forest on distances from the training points
- Buffer distances between points was calculated in meters using `raster::pointDistances`

Model Comparisons

We will be comparing the predictive accuracy of the following models:

1. Spatial Gaussian copula with ZIG marginal distributions
2. Ordinary kriging via `automap`
3. Several spatial random forests with varying $num.trees = 50, 100, 150$
4. *Zero-corrected* kriging and spatial random forests where predicted values smaller than the smallest nonzero training observation are converted to 0

Model Comparison Scenarios

We investigated three scenarios in our simulation:

1. Predicting simulated total volume with only spatial information
2. Predicting simulated hemlock volume with only spatial information
3. Predicting total volume with spatial information and annual precipitation using resampled data

We will also examine how changes in the size of the training set affect the accuracy for different methods with $n = 100, 200, 300, 500, 1000, 1200$.

- Hemlock data was of particular study interest since nearly **56%** of its original values were zeros, whereas total volume had **24.3%** zeros.

Model Comparison Metrics

We used three prediction metrics to compare the performance of the models in each scenario:

1. Root Mean Squared Prediction Error (RMSPE)

$$RMSPE = \sqrt{\frac{1}{mR}\sum_{r=1}^{R}\sum_{j=1}^{m}(\hat{y}_{j|r} - y_{j|r})^2}$$

2. Signed Relative Bias (SRB)
3. Prediction Interval Coverage (PIC_{90})

We will also examine residual plots and prediction performance for zero valued observations.

Metrics: Signed Relative Bias

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

$$\text{where } \tau = \frac{1}{mR}\sum_{r=1}^{R}\sum_{j=1}^{m}(\hat{y}_{j|r} - y_{j|r})$$

- This formula derives from the fact that MSE is equal to the bias of the estimate squared plus the variance of the estimate
- A smaller absolute value of SRB indicates smaller bias in the method with a negative value indicating underprediction and a positive value indicating overprediction.[3]

Metrics: 90% Predictive Interval Coverage

$$PIC_{90} = \frac{1}{mR} \sum_{r=1}^{R} \sum_{j=1}^{m} \mathrm{I}\left(y_{j|r} \in \hat{y}_{j|r} \pm 1.645\widehat{\mathrm{se}}(\hat{y}_{j|r})\right)$$

where $\widehat{\mathrm{se}}(\hat{y}_{j|r})$ is the standard error of all the predicted values $\hat{y}_{j|r}$ in resampled dataset r.[3]

- PIC_{90} captures the proportion of actual values for the unobserved points fall within their respective 90% prediction intervals.
- A well-calibrated model with proper assumptions should have a PIC_{90} close to 90%, but since our training and test points are spatially autocorrelated, we will examine this metric from the viewpoint of comparing models against one another.

RMSPE of Simulated Total Timber Volume

Simulated Total Volume							
n	Copula	Kriging	$RFsp_{150}$	$RFsp_{100}$	$RFsp_{50}$	$RFsp_{150}(zeros)$	Kriging (zeros)
1200	246.139	238.878	251.040	250.719	252.216	251.042	238.868
1000	264.614	248.749	256.095	256.304	257.337	256.116	248.721
500	254.933	243.617	253.945	254.267	255.217	253.957	243.604
300	264.614	248.749	256.095	256.304	257.337	256.116	248.721
200	275.154	253.674	258.074	258.246	259.417	258.113	253.628
100	298.042	268.752	266.179	266.376	267.221	266.260	268.717

Cyan indicates lowest RMSPE for sample size; gray indicates highest RMSPE.

RMSPE: Simulated Hemlock Volume

Simulated Hemlock Volume							
n	Copula	Kriging	$RFsp_{150}$	$RFsp_{100}$	$RFsp_{50}$	$RFsp_{150}(zeros)$	Kriging (zeros)
1200	48.391	46.609	48.631	48.567	48.799	48.632	46.594
1000	50.500	48.318	50.197	50.268	50.442	50.197	48.309
500	51.081	48.821	50.755	50.839	51.026	50.756	48.807
300	51.456	49.879	51.040	51.120	51.332	51.041	49.866
200	52.030	50.139	51.161	51.192	51.396	51.161	50.123
100	52.542	51.560	51.671	51.679	51.911	51.671	51.546

RMSPE: Resampled Total Volume

Resampled Total Volume							
n	Copula	Kriging	$RFsp_{150}$	$RFsp_{100}$	$RFsp_{50}$	$RFsp_{150}(zeros)$	Kriging (zeros)
1200	296.905	303.859	293.510	294.086	295.379	293.510	303.846
1000	295.311	301.473	292.139	292.557	293.580	292.139	301.461
500	303.553	304.366	296.997	297.362	298.388	296.997	304.349
300	305.409	304.526	300.984	301.412	302.469	300.984	304.504
200	308.267	304.898	303.921	304.393	305.360	303.922	304.867
100	313.791	305.210	309.662	310.072	310.971	309.669	305.159

SRB: Resampled Total Volume

Simulated Total Volume							
n	Copula	Kriging	$RFsp_{150}$	$RFsp_{100}$	$RFsp_{50}$	$RFsp_{150}(zeros)$	Kriging (zeros)
1200	-.146	-.001	.003	.003	.003	.003	-.001
1000	-.155	.001	.009	.009	.009	.009	.001
500	-.152	.001	.007	.007	.006	.006	.001
300	-.161	.002	.004	.003	.003	.003	.002
200	-.194	.000	-.001	-.001	-.001	-.002	.000
100	-.134	.006	.003	.003	.003	.001	.006

SRB: Simulated Hemlock Volume

Simulated Hemlock Volume							
n	Copula	Kriging	$RFsp_{150}$	$RFsp_{100}$	$RFsp_{50}$	$RFsp_{150}(zeros)$	Kriging (zeros)
1200	-.184	.014	.012	.011	.011	.012	.015
1000	-.190	.001	.004	.004	.004	.004	.002
500	-.190	.000	.002	.002	.001	.002	.001
300	-.190	.003	.001	.001	.001	.001	.004
200	-.189	.002	-.002	-.001	-.001	-.002	.003
100	-.182	.011	.002	.003	.003	.002	.012

SRB: Resampled Total Timber Data

Resampled Total Volume							
n	Copula	Kriging	$RFsp_{150}$	$RFsp_{100}$	$RFsp_{50}$	$RFsp_{150}(zeros)$	Kriging (zeros)
1200	-.300	-.002	.012	.012	.014	.012	-.002
1000	-.297	.002	.018	.018	.018	.018	.002
500	-.301	.000	.013	.013	.013	.013	.000
300	-.292	.000	.015	.015	.015	.015	.001
200	-.282	.000	.012	.013	.014	.012	.000
100	-.260	.007	.008	.008	.009	.008	.007

Total Volume Residual Plots

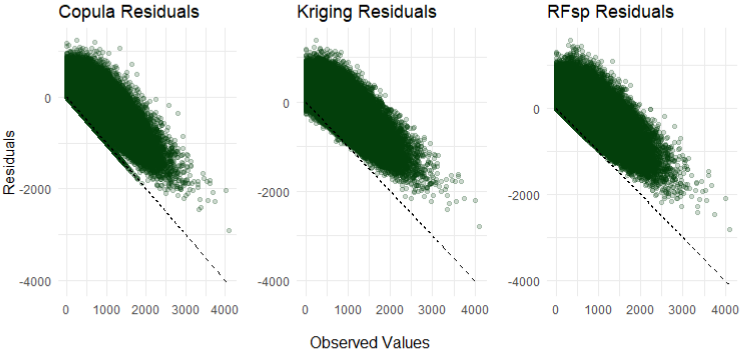

Figure: Total volume residual plots

Hemlock Residual Plots

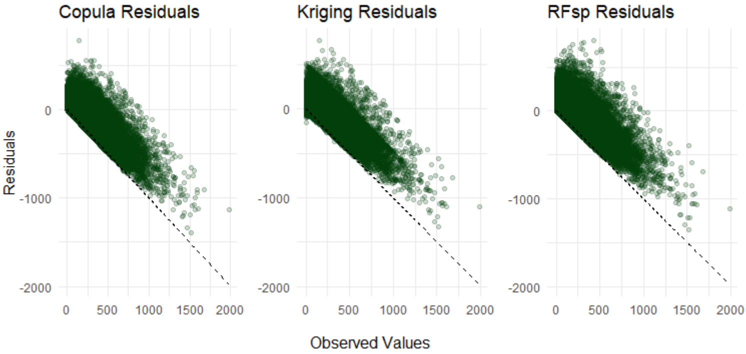

Figure: Hemlock volume residual plots

Total Volume Residual Plots (Resampled)

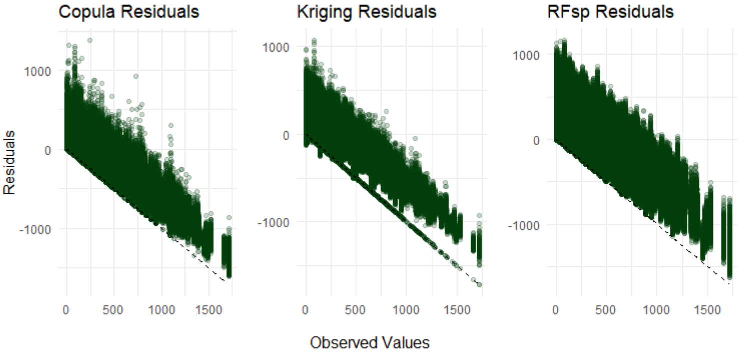

Figure: Resampled total volume residual plots

90% Prediction Interval Coverage

	Total Volume			Hemlock			Resampled		
n	Copula	Kriging	$RFsp_{150}$	Copula	Kriging	$RFsp_{150}$	Copula	Kriging	$RFsp_{150}$
1200	.841	.824	.846	.721	.569	.803	.735	.628	.795
1000	.847	.832	.855	.718	.595	.827	.739	.639	.801
500	.835	.814	.850	.700	.623	.819	.717	.629	.794
300	.812	.786	.844	.689	.640	.816	.711	.628	.785
200	.793	.759	.835	.676	.630	.803	.706	.633	.775
100	.698	.686	.809	.630	.590	.769	.705	.650	.751

Prediction of zero values

In the resampled data study with $n = 500$, we also calculated RMSPE and median predictions for points with an observed value of 0.

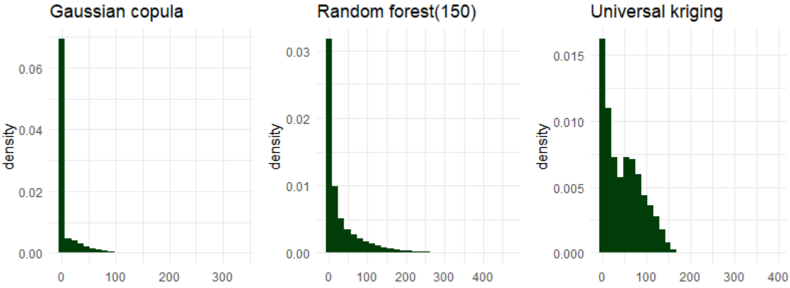

Figure: Predictions for zero values

Median $\hat{y}$			RMSPE		
Copula	$RFsp_{150}$(zeros)	Kriging(zeros)	Copula	$RFsp_{150}$(zeros)	Kriging(zeros)
0	7.49	40.5	20.7	59.6	62.4

Conclusion

- The simulations in our study only covered a small subset of forestry inventory scenarios, but with the prediction metrics we selected, kriging matched or outperformed random forests and Gaussian copula by most measures.
- While both ordinary and universal kriging had a few data artifacts in the form of negative predictions, the kriging models consistently produced unbiased estimates with relatively low RMSPE.
- Both kriging and random forest models also had low absolute values of SRB, suggesting miniscule bias, if any.

Conclusion

- In contrast, our results suggest that the Gaussian copula model underpredicts values moreso than the other two techniques, which may be due to an overabundance of zeros in the predictions.
- Given the SRB metrics for each model, we might reasonably posit that model bias played a role in inflating the copula model's RMSPE.
- However, if properly estimating unobserved points which contain zero are of significant practical importance, the Gaussian copula far outperforms both random forest and kriging.

Conclusion

- For the semicontinuous, skewed responses we simulated, every single method underestimated large values, as evidenced by the downward trending residual plots we generated for each scenario.
- The residual plots also showed that the random forest predictions had greater variance than either the copula or the kriging.
- This larger variance also mainfests itself in the PIC_{90} metrics where the random forest consistently had the greatest coverage among the methods.

References

- Hengl et. al. "Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables". In: *PeerJ - Life and Environment* (2018). DOI: 10.7717/peerj.5518.
- Noel Cressie. *Statistics for Spatial Data*. John Wiley and Sons, 1993.
- Hailemariam Temesgen Jay M. Ver Hoef. "A Comparison of the Spatial Linear Model to Nearest Neighbor (k-NN) Methods for Forestry Applications". In: *PLoS ONE* (2013). DOI: https://doi.org/10.1371/journal.pone.0059129.
- Lisa Madsen. "Maximum Likelihood Estimation of Regression Parameters with Spatially Dependent Discrete Data". In: *Journal of Agricultural, Biological, and Environmental Statistics* 14 (2009), pp. 375–391. DOI: 10.1198/jabes.2009.07116.