Comparison of Gaussian copula and random forests

in zero-inflated spatial prediction

Nick Sun

Oregon State University, Department of Statistics

May 31, 2020

Outline

1. Introduction
2. Methods
3. Results
   - RMSPE
   - SRB
   - Residual Plots
   - Prediction Interval Coverage
   - Prediction of zero values
4. Conclusion
5. References

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.

Forestry Data

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

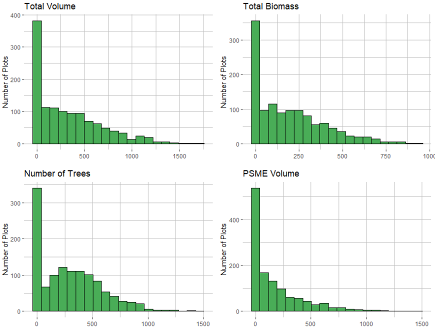

Figure: Histograms of forestry inventory variables.

Simulated Data

- 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.
- 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.

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 annual average temperature and precipitation.

Annual Precipitation and Total Timber Volume

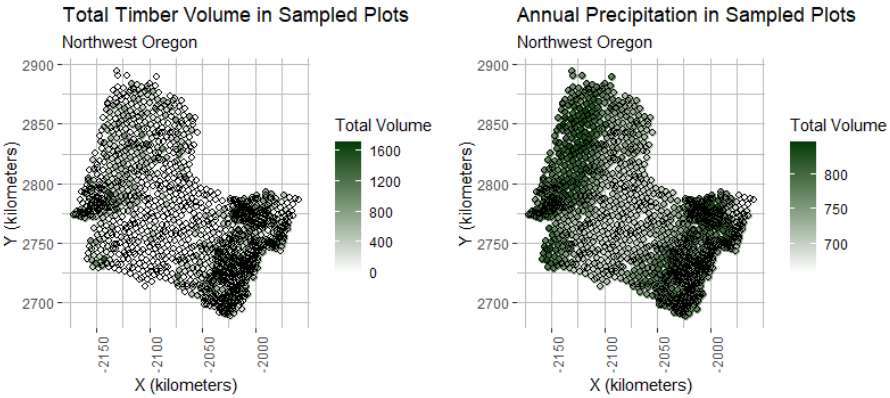

Figure: Alber's Equal Area Conic projection used here.

New Spatial Models

- These simulations will compare the predictive performance of spatial Gaussian copula, spatial random forest, and kriging in different scenarios and sample sizes.
- Two new techniques have been proposed to estimate spatially dependent data: spatial Gaussian copula and spatial random forests.

Kriging Primer

- 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*, this is, if 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*[3].

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

Kriging Primer

Kriging is often thought of as a two-step process where:

1. Spatial covariance is determined by fitting a *theoretical variogram* to the *experimental variogram*
2. Observation weights are calculated using this covariance structure and used to interpolate or predict unobserved points

Kriging Primer

- Often times, a theoretical variogram model is fit to the experimental variogram using interactive tools such as `geoR::eyefit`.
- For the purposes of this simulation study, the `automap` package will be used which relies on restricted maximum likelihood methods from the `gstat` package to fit the appropriate nugget and sill parameters, select the best theoretical model, and fit a kriging model.

Spatial Gaussian Copula

- Copulas are multivariate cumulative distribution functions where each variable has a standard uniform marginal distribution.
- An important copula result is Sklar's Theorem:

Theorem

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[5] 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

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

  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, however this method requires 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

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

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. We have chosen to use 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 necessary to make the continuous component less heavily skewed.
- Additionally, for the purposes of the copula model, zero values were instead replaced with 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 random forest is a machine learning algorithm which creates an ensemble of weak decision tree learners from bootstrapped (also referred to as *bagged*) samples of the original data[2].
- Random forest have been used in spatial prediction, but the spatial information is often disregarded[1].
- One of the notable advantages of using a machine learning algorithm like random forests is that no statistical assumptions are required, therefore, we are not required to transform the shape of the data as we had to in the Gaussian copula model.

Spatial Random Forests

- In order to incorporate this information in the model, the **RFsp** packages introduces the spatial random forest which uses buffer distances from observed points as explanatory variables.
- The generic spatial random forest system is proposed in terms of three main input components:

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

where X_G are covariates based on geographic proximity or spatial relationships, and X_R and X_P are referred to respectively as surface reflectance covariates and process-based covariates.

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 $n.trees = 50, 100, 150$
4. Semicontinuous corrected kriging and spatial random forests where predicted values smaller than the smallest nonzero training observation are converted to 0

- Testing will be done on simulated total volume and hemlock volume, as well as the resampled original data.
- Hemlock data was of particular study interest since nearly **56%** of its original values were zeros, possibly representing a more significant challenge to model than total volume which had **24.3%** zeros.
- We will also examine how changes in the size of the training set affect the accuracy for different methods with $n = 100, 200, 300, 500, 1000, 1200$.

Metrics: RMSPE

Our first prediction metric is root mean squared error

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

where lower values are preferable to higher ones

Metrics: Signed Relative Bias

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

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 mean squared prediction error 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.[4]

Metrics: 90% Predictive Interval Coverage

$$PIC_{90} = \frac{1}{mR} \sum_{r=1}^{R} \sum_{j=1}^{m} I\left(\hat{y}_{j|r} - 1.645\hat{\text{se}}(\hat{y}_{j|r}) \geq y_{j|r} \cap y_{j|r} \leq \hat{y}_{j|r} + 1.645\hat{\text{se}}(\hat{y}_{j|r})\right)$$

where $\hat{\text{se}}(\hat{y}_{j|r})$ is the standard error of all the predicted values $\hat{y}_{j|r}$ in resampled dataset r.[4]

- 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	$RFsp_{150}$	$RFsp_{100}$	$RFsp_{50}$	$RFsp_{150}(zeros)$
1200	246.139	238.878	251.040	250.719	252.216	251.042
1000	264.614	248.749	256.095	256.304	257.337	256.116
500	254.933	243.617	253.945	254.267	255.217	253.957
300	264.614	248.749	256.095	256.304	257.337	256.116
200	275.154	253.674	258.074	258.246	259.417	258.113
100	298.042	268.752	266.179	266.376	267.221	266.260

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

RMSPE: Simulated Hemlock Volume

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

RMSPE: Resampled Total Volume

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

SRB: Resampled Total Volume

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

SRB: Simulated Hemlock Volume

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

SRB: Resampled Total Timber Data

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

Residual Plots

- We produced residual plots for the Gaussian copula, kriging, and random forests with $num.trees = 150$ in each of the simulation scenarios with sample size $n = 500$.
- The dotted line on each plot corresponds indicates a predicted value of 0.
- We see that regardless of model or simulation method, $\hat{y}$ tended to underestimate large values of the observed response.

Residual Plots

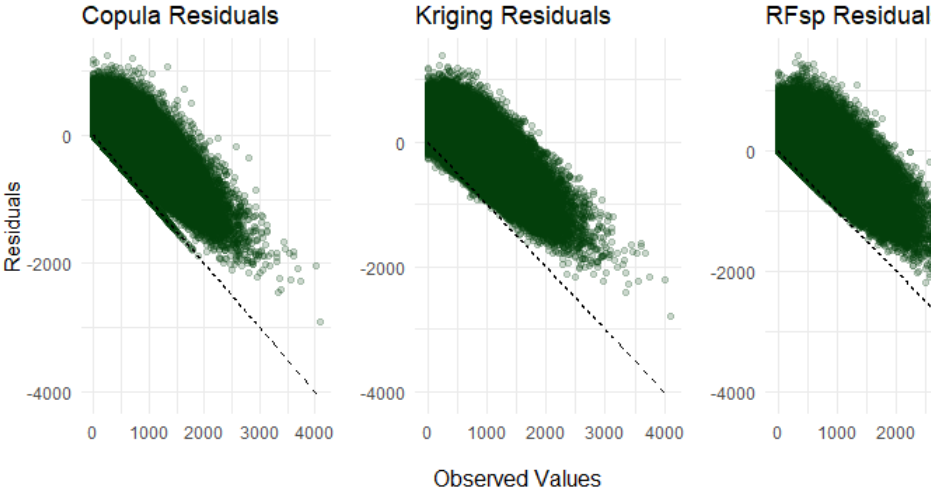

Figure: Total volume residual plots

New Spatial Models

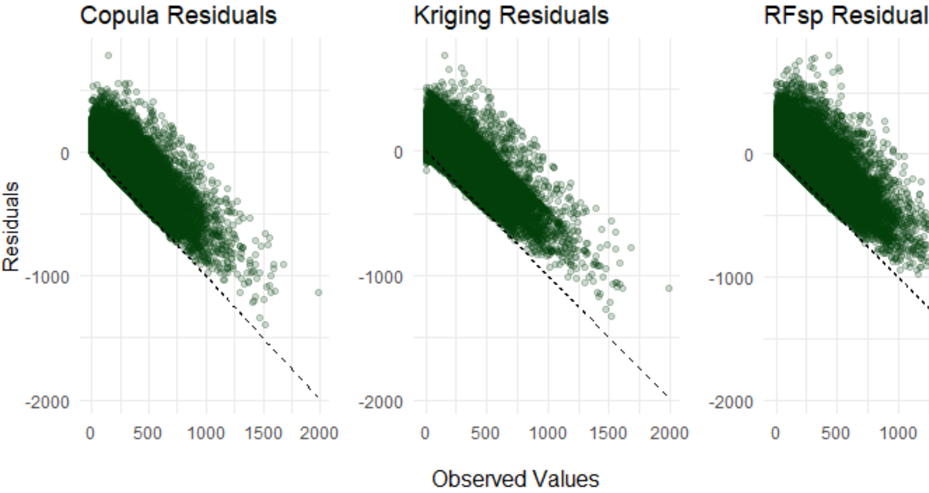

Figure: Hemlock volume residual plots

New Spatial Models

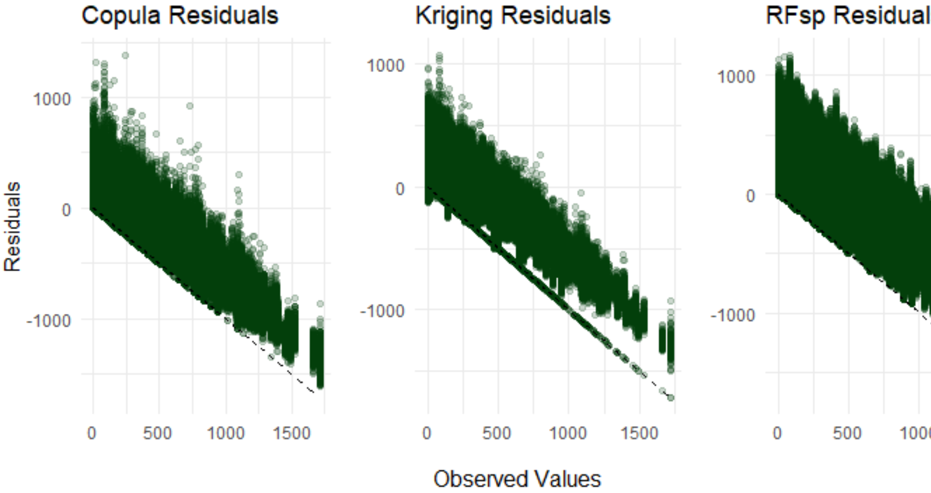

Figure: Resampled total volume residual plots

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 among the different methods for points with an observed value of 0.

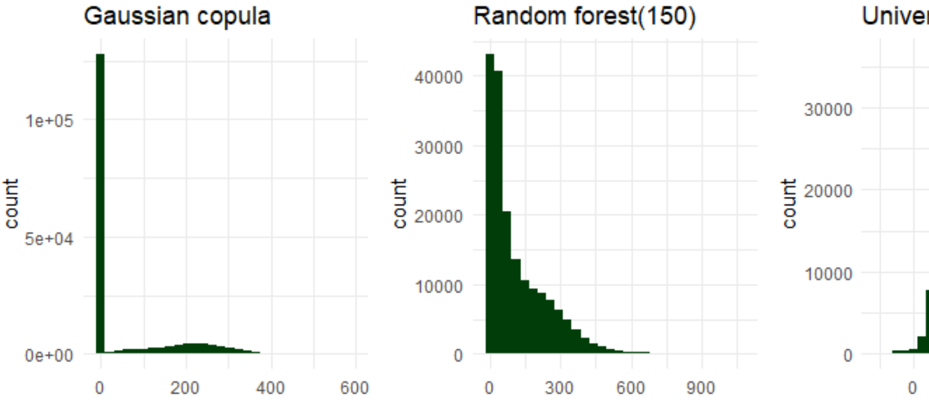

Figure: Predictions for zero values

Prediction of zero values

Figure 6 displays the histograms of the predicted values for the Gaussian copula, zero-corrected random forest, and zero-corrected universal kriging.

Median $\hat{y}$			RMSPE		
Copula	$RFsp_{150}$(zeros)	Kriging(zeros)	Copula	$RFsp_{150}$(zeros)	Krig
0	61.3	131	115	170	

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.

- Leo Breiman. “Random Forests”. In: *Machine Learning* 45 (2001).

- 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.