Comparison of Gaussian copula and random forests in zero-inflated spatial prediction for forestry applications

Nick Sun

Oregon State University, Department of Statistics

May 31, 2020

Abstract

Forestry inventory is a critical part of monitoring and servicing ecosystems and often involves statistical estimation of quantities such as total wood volume. However, forestry data is often zero-inflated, heavily skewed, and spatially dependent, making it difficult to model using traditional statistical and geostatistical models. Two new techniques have been proposed to estimate spatially dependent data: spatial Gaussian copula and spatial random forests. In this paper, we compare the predictive performance of these new models along with traditional kriging on both simulated and resampled data.

1 Introduction

An important component of forest maintenance is regular inventory of forestry resources, such as total timber volume, total biomass, etc. 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. Spatial estimation and interpolation is often employed to fill gaps in sampling and calculate estimations of relevant inventory quantities. However, forestry data has several qualities that make it difficult to model.

Forestry data taken at sampled points or plots are likely to be correlated with data points that are close by. This is what is popularly known as Tobler's First Law: "Everything is related to everything else, but near things are more related than distant things"[8]. This dependency structure precludes classical statistical models like ordinary least squares regression since those techniques rely on the assumption of independent and identially distributed data.

Furthermore, forestry data is often *semicontinuous* in that its distribution contains a point-mass at value 0 and a positive skewed distribution[9]. This overdispersion often requires modeling using a mixture distribution which combines two data generating processes: one which only generates zeros and another which generates nonnegative, continuous values. Modelling using these mixure distributions has been explored

thoroughly in non-spatial cases, but standard spatial prediction and interpolation tools such as those available in **ArcGIS Geoanalyst Toolbox**[1] do not have specialized methods to handle this semicontinuous data.

This gives need for a geostatistical model which can incorporate spatial dependence and model overdispersion of zeros. In this paper, we give a brief overview of spatial random forests from the **RFsp** `R` package[1] and the spatial Gaussian copula models[7] and compare their predictive performance in forestry applications using both simulated and resampled data.

2 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.

The response variables of interest include total volume, total biomass, total number of trees, and volume of specific tree species. Histograms of the response variables indicate that the data are positively skewed and zero inflated. Additionally, the dataset includes fuzzed[2] latitude, longitude, and elevation information. Possible covariate variables include annual precipitation, tc3 wetness index[10], annual temperature, NDVI, and cover.

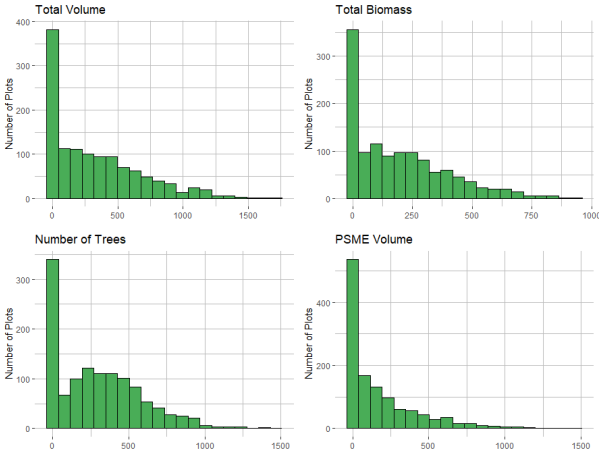

Figure 1: Histograms of forestry inventory variables.

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.

Due to the difficulty of simulating the covariates such that the original relationships between the covariates and the responses were preserved, only the response variables were generated. The models will be trained solely on the geographic locations and response values of the points in the training set. Finally, each simulated datasets is randomly split into a training data and test set.

[1] See ESRI documentation for more detail
[2] White noise is added to protect privacy of private landowners.

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. Since the focus of this study is not inference or exploration, we will focus on covariates that are known from previous work[7] to be related to forestry inventory.

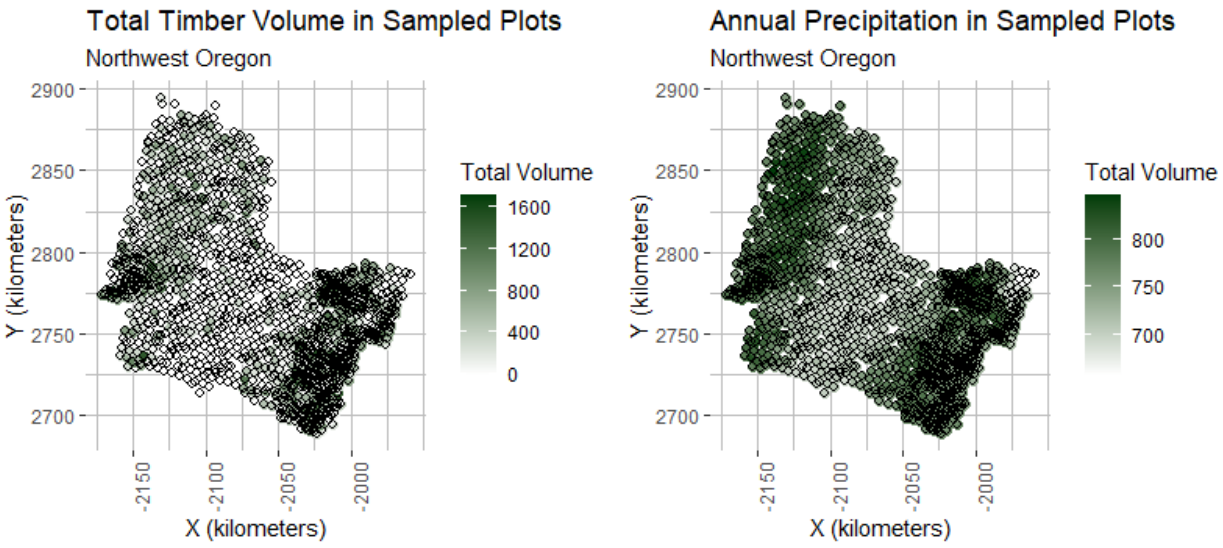

Figure 2: Alber's Equal Area Conic projection used here.

Plotting the total timber volume alongside annual precipitation reasonably suggests that timber is associated with level of precipitation. Figure 3 is the semivariogram with the annual precipitation effect incorporated, indicating that spatial autocorrelation effects are greatly reduced. This will be valuable in determining model performance when spatial correlation effects are minimal and auxiliary covariates are incorporated.

3 Methods

These simulations will compare the predictive performance of spatial Gaussian copula, spatial random forest, and kriging in different scenarios and sample sizes.

3.1 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 many regards, kriging is very similar in principle to

regression analysis. 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]. The weights w obtained by kriging are unbiased and minimize estimation variance.

$$\hat{z}(x_0) = \sum_{i=1}^{N} w_i z(x_i)$$

where x_0 is the point to be predicted and $z(x_i)$ are the observed points.

If the response at point u_α is defined as the function $Z(u_\alpha)$, covariance between points is estimated using the sample semivariogram which is defined for a lag distance h as

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (Z(u_\alpha) - Z(u_\alpha + h))^2$$

where $N(h)$ is the number of pairs separated by distance h. An underlying theoretical population variogram model is then fit to $\hat{\gamma}(h)$, such as the Gaussian variogram model:

$$\gamma(h) = c_0 + c_1 \left(1 - \exp\left(-\frac{h}{a}\right)^2\right)$$

Using this theoretical variogram function, we can calculate the covariance $C(h) = \sigma^2 - \gamma(h)$ for any lag distance h where σ^2 is the sample variance of all points. 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

The original timber volume data was found to fit best with a Gaussian model, however, the Gaussian model may not be the best fit for the training datasets we created, particularly in the resampling data study.[3] 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.

[3]Several of our simulated datasets were also found to fit best with an exponential or spherical variogram model.

As an estimation approach, kriging makes use of distance between points as well as axes of spatial continuity and redundancy of data points. Kriging therefore is a very popular technique among spatial analysts since it incorporates a lot of information into the modelling process. However, kriging still has underlying assumptions of a Gaussian process, potentially making it ill-suited for semicontinuous data.

Ordinary and Universal Kriging A common point of confusion for newcomers of geostatistics is that kriging can refer to a variety of related spatial interpolation techniques with different nomenclature. For consistency with the `gstat` and `automap` packages, we will refer to the two kriging techniques we use in this paper as *ordinary kriging* (OK) and *universal kriging* (UK). Both techniques rely heavily on the process outlined above and therefore are very similar in principle.

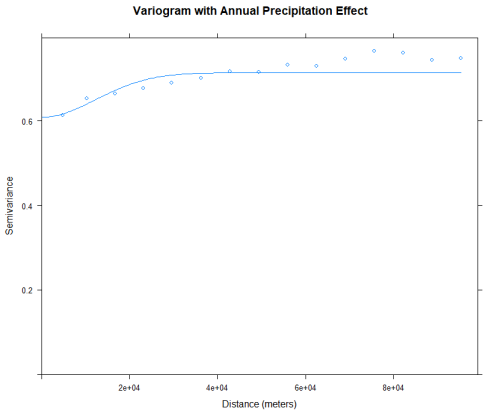

Figure 3: Note that there is almost no change in semivariance as distance increases.

Ordinary kriging is used for simulations that do not involve any covariate. For the scope of this work, this means that ordinary kriging is used for studying our simulated datasets, but not our resampled ones. More technically, OK assumes a constant unknown mean in the local neighborhood of each estimation point.

This differs from universal kriging (referred to as *regression kriging* or *kriging with external drift* in other sources) which assumes an overall smooth, nonstationary trend in the data which can be described as a function of auxiliary predictors and a random residual which is estimated from residuals of the observed points.[6] Universal kriging is used in simulations which involve using annual precipitation as a covariate.

3.2 Spatial Gaussian Copula

Copulas are multivariate cumulative distribution functions where each variable has a standard uniform marginal distribution. Copulas were developed to describe dependency structures between random variables and have been previously applied in areas such as microRNA[11] and box-office data[5]. An important copula result is Sklar's Theorem states that every 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

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

There are many possible choices for C, but a popular selection is the multivariate normal CDF Φ_Σ where Σ is the correlation matrix describing the relationship between the variables.

Madsen[7] 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, however this method requires the appropriate selection of F and Σ.

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. These parameters can be estimated from the original data[7].

An appropriate F function would be one which can handle semicontinuous data. In this paper, we have chosen to use a zero-inflated gamma (ZIG) 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 makes 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.

The complete spatial Gaussian copula algorithm used here is detailed below:

Algorithm 1: Spatial Gaussian Copula

Result: Predictions for unobserved locations
for *Each simulated dataset* **do**
  Cube root transform observed responses;
  Find smallest nonzero responses ϵ ;
  Transform 0s into small $U(0, \epsilon)$ random variables;
  Calculate spatial covariance parameters θ_N, θ_R and ZIG parameters β, π;
  if *covariates present* **then**
    calculate β, π using logistic and Gamma GLM with the covariates;
  else
    calculate β, π using logistic and Gamma intercept-only GLM;
  end
  Transform responses to standard uniform using CDF of zero-inflated Gamma;
  Use kriging on the standard normal random variables to get estimates for the unobserved values ;
  Backtransform unobserved standard normal values to get predictions for the unobserved values on the original scale;
end

3.3 Spatial Random Forest

The random forest is a machine learning algorithm which creates an ensemble of decision trees from bootstrapped (also referred to as *bagged*) samples of the original data[2]. Each of the n decision trees is trained on a random subset of variables at each split in the tree. While individual decision trees are prone to overfitting on training data, a large collection of randomly generated weak learners is less prone to these biases. The prediction of the random forest is taken as the mode or average of the entire ensemble. 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.

Random forests have been used in spatial prediction, but the spatial information is often disregarded[1]. Ignoring spatial autocorrelation can result in biased predictions. 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. Common examples of surface reflectance covariates would be spectral bands from remote sensing images. Process-based covariates are more traditional independent variables, for example, average annual precipitation. Not all types of covariates need be present to create a spatial random forest and previous work by Hengl et. al. has demonstrated that

including only X_G generates predictions similar to ordinary kriging while including X_G and X_P generates predictions similar to universal kriging[1].

The **RFsp** packages is built on top of the **ranger** `R` package which supports high dimensional datasets. However, the authors of spatial random forest caution that since distances need to be calculated in order to include spatial information, **RFsp** might be slow for large datasets.

Algorithm 2: Spatial Random Forest

Result: Predictions for unobserved locations
for *Each simulated dataset* **do**
  The buffer distances between each point in the training set is calculated;
  n random samples are drawn with replacement from the training data;
  n trees are generated from the random samples with the buffer distances as covariates;
  The buffer distances between each unobserved location and the points in the training set is calculated;
  These buffer distances are input into the random forest and a prediction is generated;
end

4 Results

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. Semicontinuous *zero-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 $n = 100, 200, 300, 500, 1000, 1200$. The metric of interest will be root mean square prediction error:

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

where $r \in R$ is a simulated dataset and $j|r$ signifies the prediction for observation j in the dataset r.

We will also examine the *signed relative bias* of each pointwise prediction method:

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

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]

Our final prediction metric is 90% *prediction interval coverage* for each of the methods:

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

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.

In addition to these metrics, we will also examine the residuals and specific prediction performance of zero values in each model.

4.1 RMSPE

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.

For small sample sizes in the total volume simulation, the copula model had between 5% and 10% higher RMSPE than the kriging or random forest models. As n increases, the copula model had lower RMSPE than the random forests. Kriging and zero-corrected kriging had the lowest RMSPE for most sample sizes, narrowly outperforming random forests.

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

We see a similar pattern to the hemlock volume simulation where kriging and zero-corrected kriging had the lowest RMSPE. The copula model had the highest RMSPE except for $n = 1200$, however the relative differences between all the models are smaller than in the total volume simulation.

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

In the total volume resampling study, random forests with $num.trees = 150$ have the best RMSPE across the board. For small training sizes, the copula model has the highest RMSPE but this changes when n grows large where it outperforms kriging and zero-corrected kriging.

4.2 Signed Relative Bias

In all simulations and sample sizes, the copula model showed negative bias indicating that predictions tend to be underestimated. Random forests and kriging models show minimal bias.

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

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

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

4.3 Residual Analysis

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.

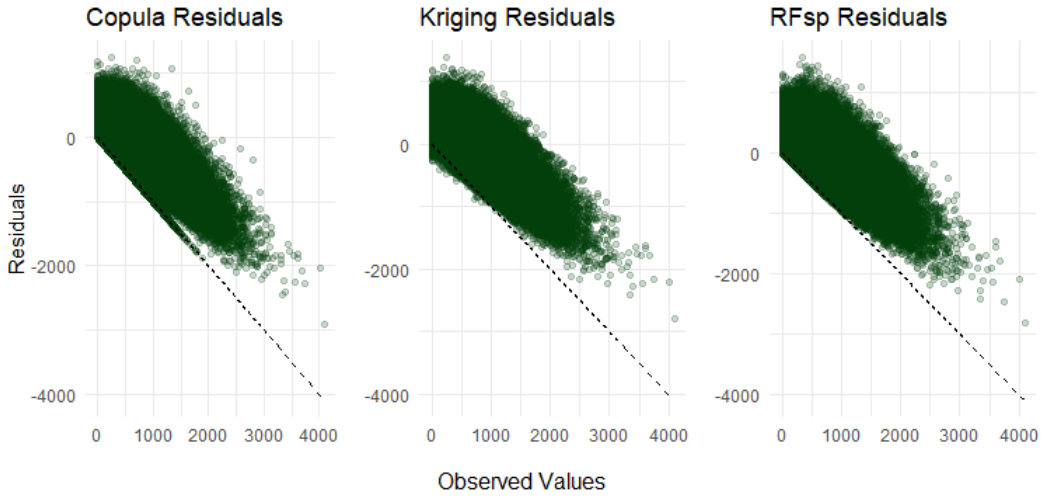

Figure 4: Total volume residual plots

The residual plots in figure 4 are similar for the most part. There is a distinct line for the zero predicted values in the copula model whereas the kriging model predicted some negative values.

Figure 5 suggests that random forests produce residuals with higher variance than either the other models. While copula and kriging produced residual plots with almost rectangular shapes, random forests produced residuals that looked more like a cloud of points, particularly with observed values around 500.

Figure 6 shows that random forests have the widest residual spread which is relatively homoskedastic as observed values increase. This contrasts with kriging residuals which appears as a narrowing funnel as observed values increases. The kriging residuals also produced some predictions near zero, even for large values. This manifests visually as a "gap" within the kriging residuals.

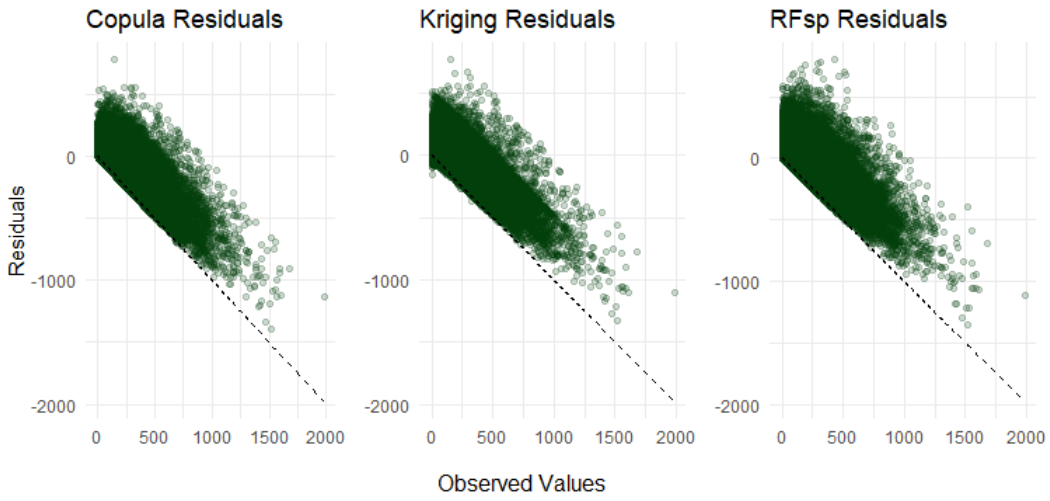

Figure 5: Hemlock volume residual plots

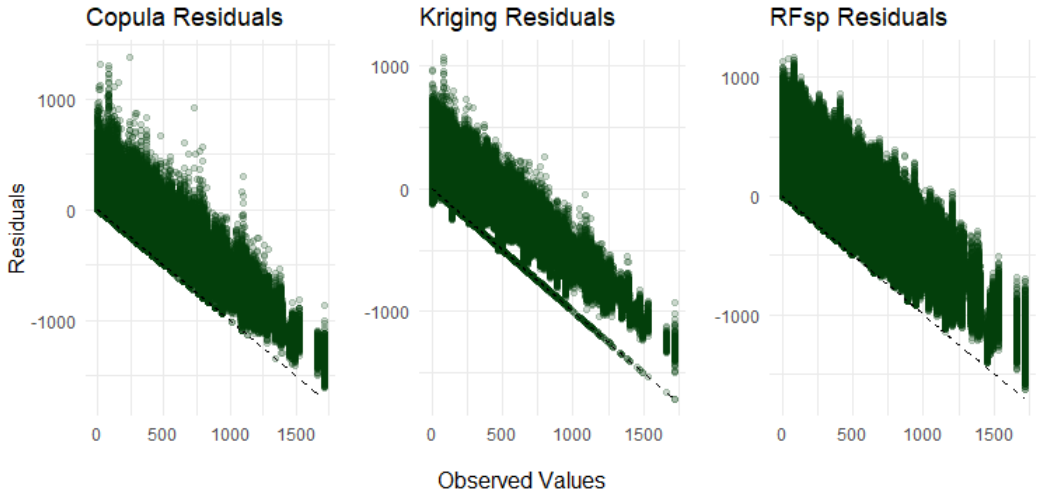

Figure 6: Resampled total volume residual plots

4.4 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

We computed PIC_{90} is computed for the Gaussian copula, kriging, and random forests with $num.trees = 150$. All of the models failed to reach 90% prediction coverage, but random forests had the greatest coverage which was fairly consistent among the different sample sizes. The copula model started with prediction

coverage below 70%, but as sample size increased both closed the gap with the random forest. Kriging had the lowest PIC_{90} across the board, particularly in the hemlock and resampled study.

4.5 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 7 displays the histograms of the predicted values for the Gaussian copula, zero-corrected random forest, and zero-corrected universal kriging.

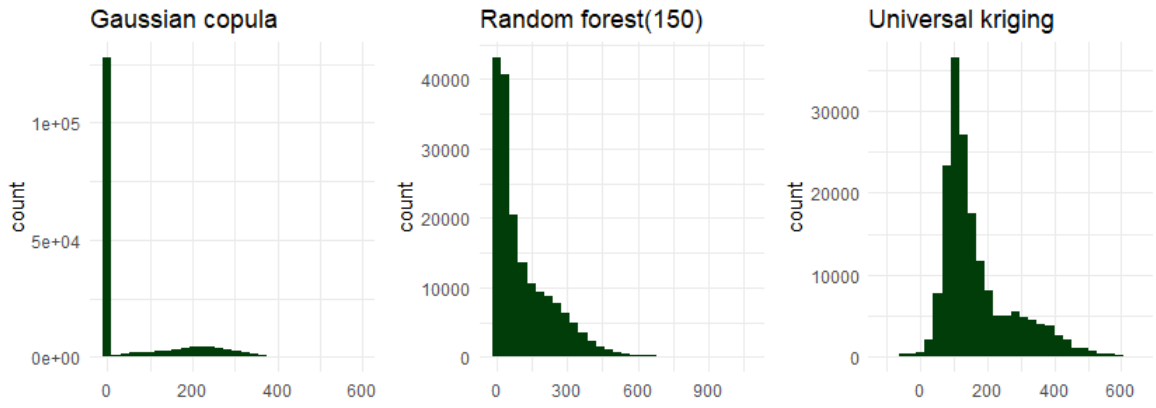

Figure 7: Predictions for zero values

Median $\hat{y}$			RMSPE		
Copula	$RFsp_{150}$(zeros)	Kriging(zeros)	Copula	$RFsp_{150}$(zeros)	Kriging(zeros)
0	61.3	131	115	170	201

The copula model far outperforms both the random forest and universal kriging model for predicting zero values. Across 1000 resampled datasets, the copula correctly predicted 72.5% of observed zero values, whereas the random forest only correctly predicted .4% of the values and universal kriging made no predictions of exactly zero.

5 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.

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.

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. High PIC_{90} might be desireable in cases where interval estimates are preferred to point estimates.

References

[1] 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.

[2] Leo Breiman. "Random Forests". In: *Machine Learning* 45 (2001).

[3] Noel Cressie. *Statistics for Spatial Data*. John Wiley and Sons, 1993.

[4] 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.

[5] Ting Liu Junwen Duan Xiao Ding. "A Gaussian copula regression model for movie box-office revenue prediction". In: *Science China* 60 (2017). DOI: 10.1007/s11432-015-0905-6.

[6] Ivana Mesic Kis. "Comparison of Ordinary and Universal Kriging interpolation techniques on a depth variable (a case of linear spatial trend), case study of the Sandrovac Field". In: *The Mining-Geology-Petroleum Engineering Bulletin* (2015), pp. 41–58.

[7] 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.

[8] Harvey J. Miller. "Tobler's First Law and Spatial Analysis". In: *Annals of the Association of American Geographers* 94 (2004), pp. 284–289. DOI: www.jstor.org/stable/3693985.

[9] Elizabeth Dastrup Mills. "Adjusting for covariates in zero-inflated gamma and zero-inflated log-normal models for semicontinuous data". PhD thesis. University of Iowa, Department of Biostatistics, 2013.

[10] Martha Raynolds and Donald Walker. "Increased wetness counfounds Landsat-derived NDVI trends in the centeral Alaska Slope region, 1985-2011". In: *Environmental Research Letters* 11 (2016). DOI: https://iopscience.iop.org/article/10.1088/1748-9326/11/8/085004.

[11] Grace Yoon, Raymond J. Carroll, and Irina Gaynanova. *Sparse semiparametric canonical correlation analysis for data of mixed types*. 2018. arXiv: 1807.05274 [stat.ME].