Comparison of Gaussian copula and random forest in zero-inflated spatial prediction

Nick Sun

May 19, 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 them difficult to model using traditional statistical and geostatistical models. Two new techniques have been proposed to estimate spatially dependent data: a spatial Gaussian copula model and spatial random forests. In this paper, we compare the predictive performance of these new models along with ordinary 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.

Simply put, forestry data at sampled points or plots is 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"[7]. This dependence 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[8]. 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. 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[6] 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.

[1]See ESRI documentation for more detail

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[9], annual temperature, NDVI, and cover.

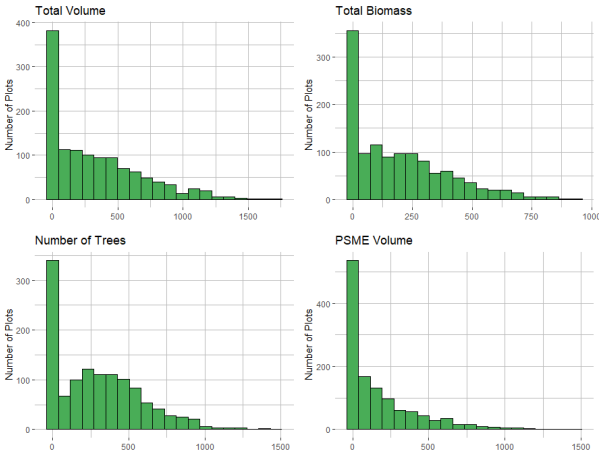


Figure 1: Histograms of forestry inventory variables.

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. After generation, we randomly sample points from each of the datasets to use as training data for the copula, random forest, and kriging models. Afterwards, the models predictive capabilities are then tested on the remaining points.

It was difficult to also simulate the covariates which preserve the original relationships between the covariates and the responses, therefore, only the response variables of interest were simulated for this study. The models will be trained solely using the geographic locations of the data points and the values at each of the simulated points. We simulated $m = 1000$ datasets of size $n = 1224$ for two different variables: total timber volume and hemlock volume. Total timber volume is a common practical variable of interest in forestry inventory applications. Hemlock data is also of 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.

Resampled Data Using the original Oregon dataset, we generate datasets of various sizes by sampling rows without replacement. We treat these sampled rows as the training data for our models and the remaining points as unobserved test points. For these simulations, 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 data exploration, we will focus on covariates that are known from previous work[6] to be related to forestry inventory.

Plotting the total timber volume alongside annual precipitation reasonly suggests that timber is associated with level of precipitation, although high precipitation does not always mean high timber volume, as seen in the many sampled points along the western edge of the study area. From the semivariogram with the annual precipitation effect incorporated, we see from the ratio of the sill to nugget effect 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.

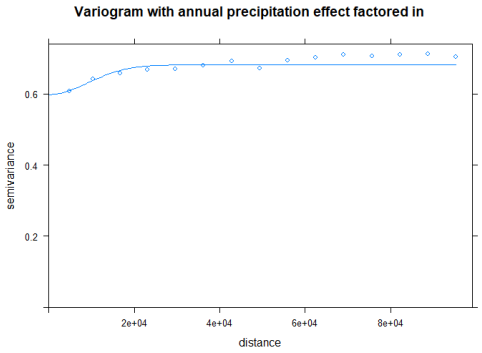


Figure 3: Note that there is almost no change in semivariance as distance increases.

3 Methods

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

[2]White noise is added to protect privacy of private landowners.

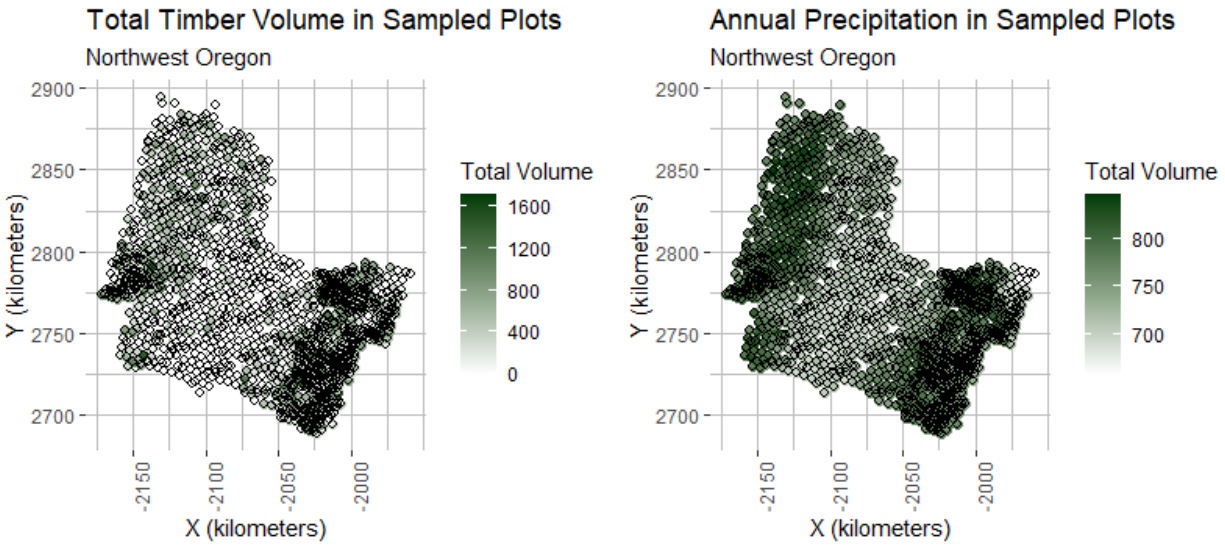


Figure 2: Alber's Equal Area Conic projection used here.

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* via the Gauss-Markov theorem. The weights w obtained by kriging are unbiased and minimize estimation variance.

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

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

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

where $N(h)$ is the number of pairs separated by distance h. This function $\gamma(h)$ is one half the average of squared differences on the pairs $N(h)$ and quantifies the relationship of difference over distance. Using this function, we can calculate the covariance $C(h) = \sigma^2 - \gamma(h)$ for any lag distnace 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

An example of an experimental variogram with an overlaid theoretical model is shown. There are several common choices for theoretical semivariogram models: spherical, exponential, Gaussian, etc. The original timber volume data was found to fit best with a Gaussian model which has a sigmoidal shape. However, the Gaussian model may not be the best fit for the datasets we simulate[3], particularly in the resampling data study.

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

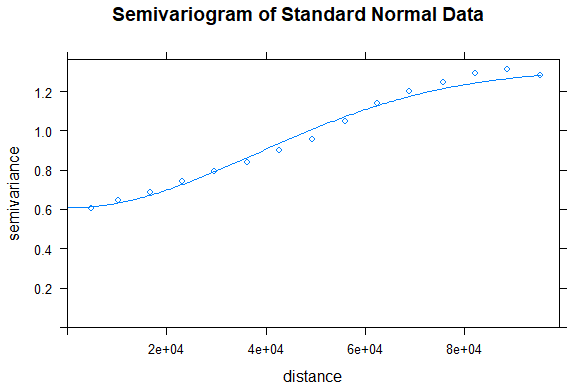


Figure 4: An example of a Gaussian variogram

Often times, a theoretical variogram model is fit to the experimental variogram using interactive tools such as `geoR::eyefit` or using maximum likelihood methods. 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.

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.

Ordinary kriging is used for simulations that do not involve a covariate, which in our scenario will be annual precipitation. 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.[5] 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 to microRNA[10] and box-office data[4]. 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 $C : [0,1]^n \to [0,1]$.

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

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[6] 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. Differentiation 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.

An appropriate F function would be one which can handle semicontinuous data. In this paper, 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.

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 weak decision tree learners 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 forest 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. Commons 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

For this simulation study, 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 small values are converted to 0

We will also examine how changes in the size of the training set affect the accuracy for different methods $n = 100, 200, 300$. The metric of interest will be root mean square prediction error (RMSPE), defined as

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

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 using the following formula[3]

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

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. While bias itself does not tell us much, bias as a ratio of MSE allows us to investigate the error for different prediction methods deeper than MSE alone. A smaller absolute value of SRB means smaller bias in the method with a negative value indicating underprediction and a positive value indicating overprediction.

Lastly, we will examine the 90% *prediction interval coverage* for each of the methods, defined as

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

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.[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.

4.1 Simulation Results - Total Volume

4.1.1 RMSPE

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

For most sample sizes, the copula model had between 5% and 10% higher RMSPE than the kriging or random forest models, although these results suggest that the difference grows smaller as n increases and for large training sets, the copula model had lower RMSPE than the random forests. Kriging had the best prediction performance for most sample sizes except for $n = 100$, but random forests usually had relatively close error metrics.

4.1.2 Signed Relative Bias

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

The copula model seems to significantly underestimate values compared to random forests and kriging. Both random forest and kriging show little to no bias.

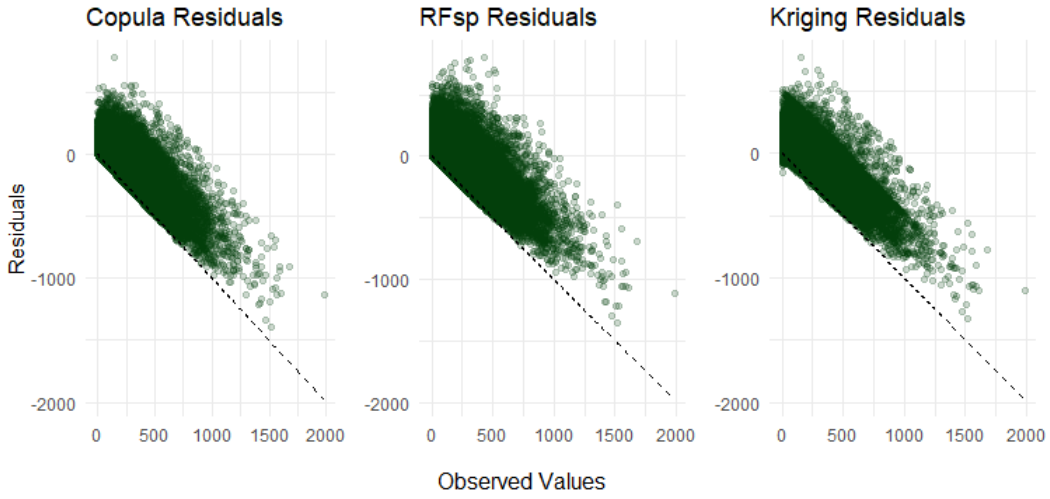
4.1.3 Predictive Interval Coverage 90%

n	Copula	Kriging	$RFsp_{150}$
1200	.841	.824	.846
1000	.847	.832	.855
500	.835	.814	.850
300	.812	.786	.844
200	.793	.759	.835
100	.698	.686	.809

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, even for the larger training sets. Random forest had the greatest coverage which was fairly consistent among the different sample sizes. Both kriging and copula models started with prediction coverage below 70%, but as sample size increased both models closed the gap with the random forest.

4.1.4 Residual analysis

We produced residual plots for the Gaussian copula, kriging, and random forests with $num.trees = 150$. The dotted line on each plot corresponds to the negative of the observed value, meaning the predicted value was 0.



We see that regardless of prediction method, $\hat{y}$ tended to be underestimated with this effect being more pronounced for extreme values of observed total timber volume. A distinct line for the zero predicted values is visible in the copula model whereas the ordinary kriging residuals show that some negative predictions were made.

4.2 Simulation Study - Hemlock Volume

4.2.1 RMSPE

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 total volume simulation where kriging had the lowest RMSPE, but the relative gap between the models is smaller. Again, the random forests have lower RMPSE than copula model, except for cases with large training set size.

4.2.2 Signed Relative Bias

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

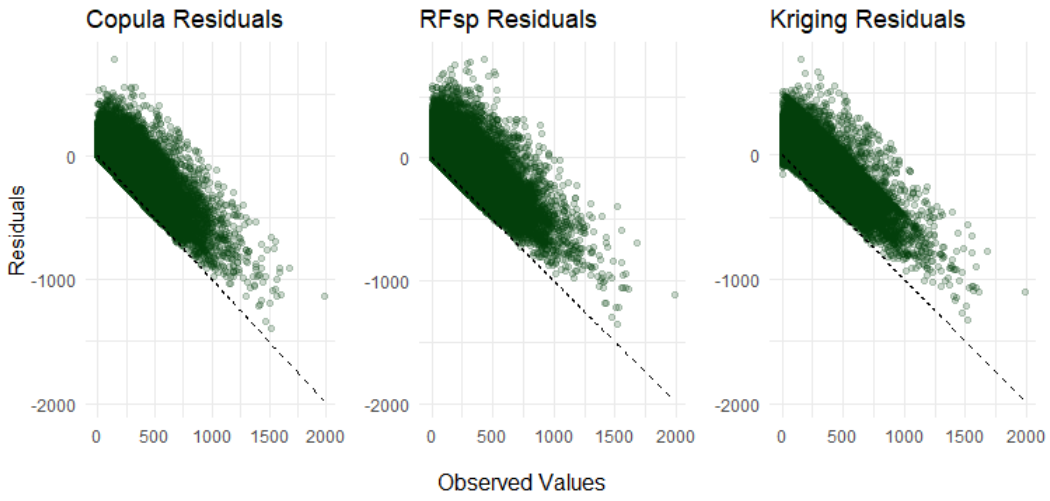
The copula model once again shows negative bias whereas the other models show minimal bias.

4.2.3 Predictive Interval Coverage 90%

n	Copula	Kriging	$RFsp_{150}$
1200	.721	.569	.803
1000	.718	.595	.827
500	.700	.623	.819
300	.689	.640	.816
200	.676	.630	.803
100	.630	.590	.769

Again, the prediction intervals do not meet 90% coverage, but the random forest does have the highest coverage and the copula model having higher coverage than the ordinary kriging model.

4.2.4 Residual analysis



The residual plots indicate that the random forest model certainly has the largest spread of predicted values compared to the copula model and the kriging model. Again, the kriging model predicted some negative values, as is visible in the residuals.

4.3 Simulation Results - Resampled Original Data

4.3.1 RMSPE

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

Interestingly, universal kriging has the highest RMSPE in the sample sizes ≥ 500.

4.3.2 Signed Relative Bias

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

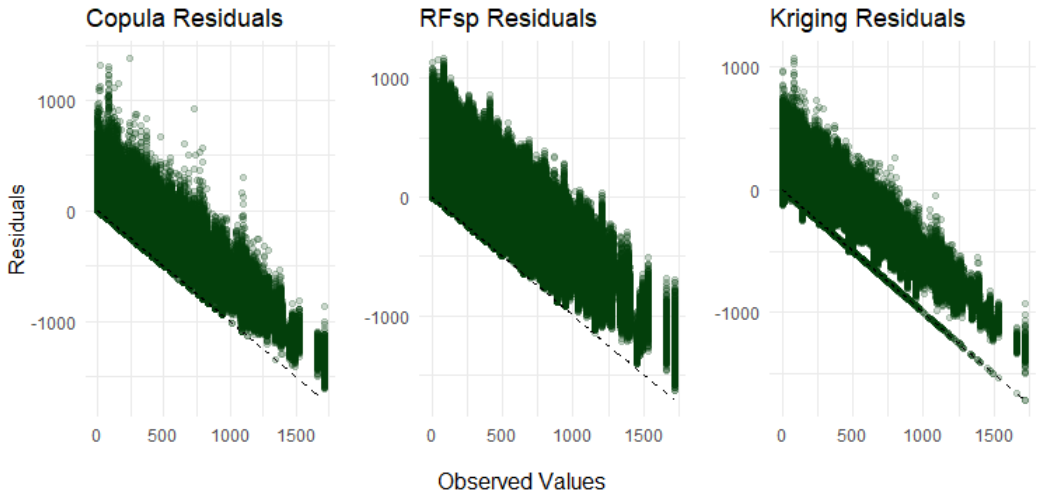
Again, we see significant negative bias in the copula model whereas the other models have little to no bias.

4.3.3 Predictive Interval Coverage 90%

n	Copula	Kriging	$RFsp_{150}$
1200	.735	.628	.795
1000	.739	.639	.801
500	.717	.629	.794
300	.711	.628	.785
200	.706	.633	.775
100	.705	.650	.751

In a similar trend to the other simulations, the random forest model had the highest coverage and the universal kriging model had the lowest.

4.3.4 Residual analysis



Again, we see the that the random forest has the widest spread of predicted values compared to the other models. Universal kriging appears to have a funnel shape towards the large observed values where the variation of predictions narrows. The copula residuals suggests that certain observed values produced a right skewed distriubtion with extreme predictions.

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. All of the 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. While there were simulated scenarios where the Gaussian copula had lower RMSPE than the random forest, particularly when the training data sets were large and only a few unobserved points were being predicted, in most scenarios random forest and kriging outperformed it. This was particularly apparent when training set size was small. Given the SRB metrics for each model, we might reasonably posit that model bias played a role in inflating the copula model's RMSPE.

For both kriging and random forests, converting small predicted values to zeros using the smallest nonzero training value as a threshold only marginally affected the prediction metrics. This makes logical sense as most cases of failing to correctly estimate a zero point was because a model instead estimated a relatively small nonzero value. These nonzero prediction would result in a relatively small residual which would likely barely affect metrics like RMSPE. However, this illustrates some possible shortcomings with random forests and kriging which are not captured in RMSPE. If properly estimating unobserved points which contain zero are of practical importance, the Gaussian copula far outperforms both random forest and kriging. The small RMSPE changes between the kriging and random forests estimates and the zero-corrected kriging and random forests estimates suggests that very few estimates were actually estimated to be precisely zero. Depending upon the forestry inventory application of interest, this might be a serious flaw with the kriging and random forest models.

It is notable that for the semicontinuous, skewed responses we simulated, every single method underestimated large values. This effect grew more pronounced as the values became larger, as evidenced by the downward trending residual plots we generated for each scenario. The residual plots also revealed that the random forest residuals had greater variance than either the copula or the kriging residuals, as indicated by the wider band of residuals. This is most clearly seen in the resampling data study. This larger variance also

mainfests itself in the PIC_{90} metrics where the random forest consistently had the greatest coverage among the methods. While none of the models in any of the scenarios we tested reached exactly 90% coverage, high PIC_{90} might be desireable in cases where interval estimates are preferred to point estimates. In other applications, the wide spread of predictions for each observed point might make the random forest model less desirable.

In closing, ordinary and universal kriging are still appear to be viable models in semicontinuous contexts, regularly outperforming both the copula and random forest in RMSPE. However, both Gaussian copula and random forests still have their use cases. The appeal of the random forest lies in the lack of statistical assumptions, making it a very flexible spatial prediction technique. Additionally, there seemed to be little differences in our simulation between random forests with different numbers of trees, suggesting that for some forestry applications, the ensemble learner does not have to be particularly large in order to get good estimates. While the authors of `RFsp` caution that the package can run slow with large datasets, in our simulations even with the largest training set of 1200 points, there was little practical difference in runtime between the random forests and the other models.

The Gaussian copula by contrast requires more statistical legwork. The marginal distributions of the points need to be known beforehand, requiring significant data exploration. Additionally, the Gaussian copula algorithm we used produced biased results and overpredicted the amount of zeros in the test data. However, in certain cases where the training data is large, the Gaussian copula did produce lower RMSPE than the random forests. The Gaussian copula was also the only model to consistently predict zero values, making it useful for applications where estimating whether or not a point has a zero value

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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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