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

Nick Sun

May 13, 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"[6]. 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[7]. 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 spatial Gaussian copula models[5] 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. Additionally, the dataset includes fuzzed[2] latitude, longitude, and elevation information. possible covariate variables include annual precipitation, tc3 wetness index[8], annual temperature, NDVI, and cover.

Histograms of the response variables indicate that the data are positively skewed and zero inflated. Furthermore, some data exploration shows that a plot of the sampled points with an overlaid of contour mapping of annual precipation amount suggests that more timber is found in areas with high precipitation, although this is not always the case as there are high-valued sampled points which fall outside the areas of high precipitation.

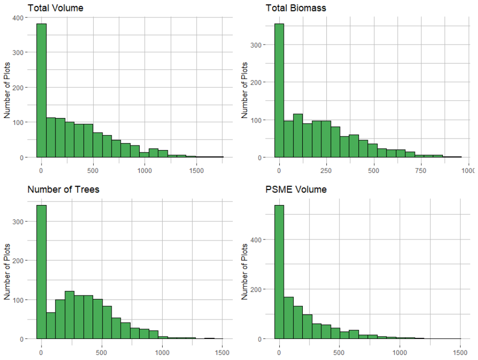
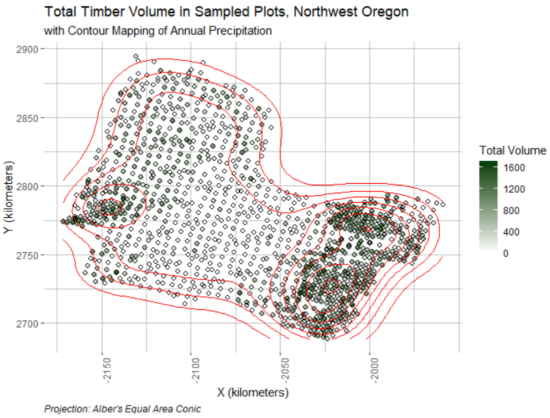


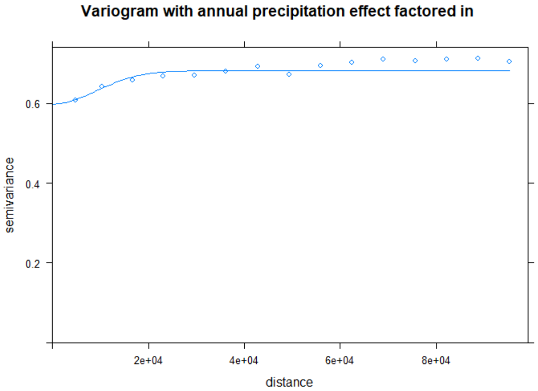
Figure 1: Histograms of response variables.

Simulated Data Using this original forestry dataset, we create $m_1 = \mathit{1000}$ simulated datasets of size $n = \mathit{1224}$ by drawing multivariate random normal samples with the calculated sample correlation matrix and backtransforming using the quantile function of the zero-inflated gamma function. We then randomly sample 300 points from each of the simulated datasets to use as training data for the copula and random forest models. Afterwards, the models predictive capabilities are tested on the remaining 924 test points.

It was difficult to also simulate the covariates which preserve the original relationships between the covariates and the responses, therefore, only the responses were simulated for this study. The Gaussian copula and the random forests will be used solely then on the geographic locations of the data points and the values at each of the simulated points.

Resampled Data Again using the original forestry dataset, we obtain $m_2 = \mathit{1000}$ resampled datasets of size $n = \mathit{300}$ by sampling rows without replacement from the data and treating that as the training data. We then compare the performance of the trained models in predicting the values of the remaining 1024 points.


Total Timber Volume in Sampled Plots, Northwest Oregon
with Contour Mapping of Annual Precipitation
2900
2850
2800
2750
2700
Y (kilometers)
-2150
-2100
-2050
-2000
X (kilometers)
Total Volume
1600
1200
800
400
0
Projection: Alber's Equal Area Conic



The contour map gives a clear graphical suggestion that total timber volume is associated with the annual precipitation. 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.

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

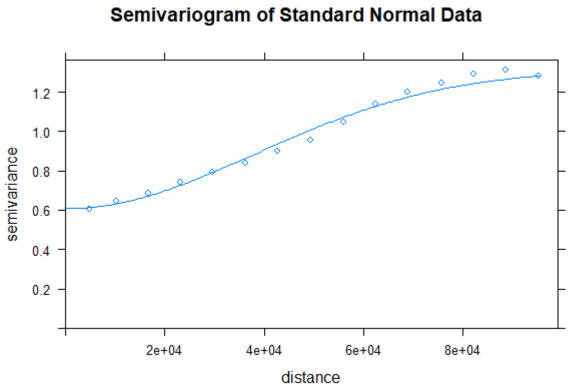


Figure 2: An example of a Gaussian variogram

3 Methods

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.

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

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.

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.

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[9] 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[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. Differentiation 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.

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. 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.

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

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 $\epsilon$ ;
      Transform 0s into small $U(0, \epsilon)$ random variables;
      Calculate spatial covariance parameters $\theta_N, \theta_R$ and ZIG parameters $\beta, \Pi$;
      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, X_R is surface reflectance covariates, and X_P are process-based covariates.

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

$$PIC90 = \frac{1}{mR}\sum_{r=1}^{R}\sum_{j=1}^{m} I\left(\hat{y}_{j|r} - 1.645\hat{\text{se}}(\hat{y_{j|r}}) < y_{j|r} \cap \hat{y}_{j|r} + 1.645\hat{\text{se}}(\hat{y_{j|r}})\right)$$

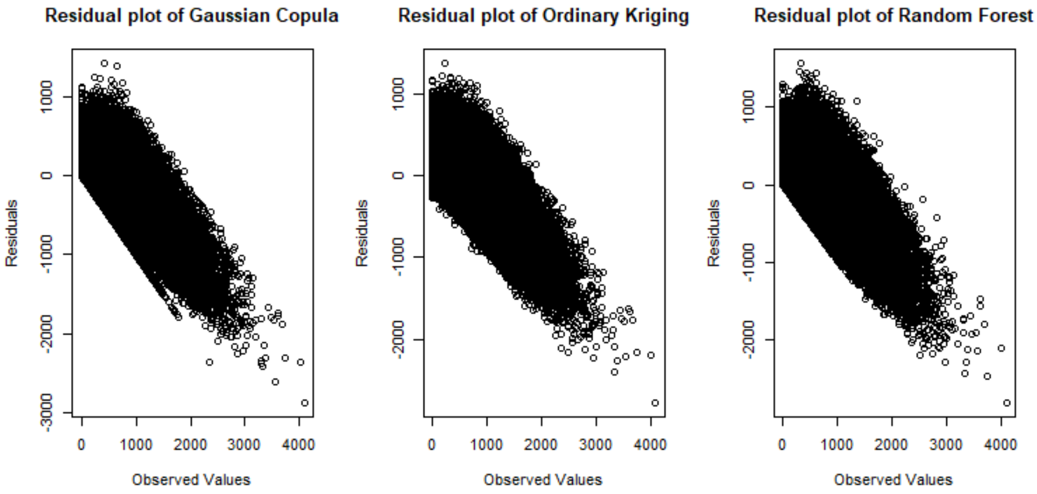
4.1 RMSPE

n	Copula	Kriging	$RFsp_{150}$	$RFsp_{100}$	$RFsp_{50}$	$RFsp_{150}(zeros)$	Kriging (zeros)
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

Our 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. Kriging had the best prediction performance for most sample sizes except for $n = 100$, but random forests usually had relatively close error metrics.

4.2 Residual analysis

We can plot the prediction error from all 1000 simulations.



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.

5 Conclusion

Here is a conclusion where I will probably say something like each technique has its own advantages and disadvantages, etc.

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] 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.

[6] 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.

[7] 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.

[8] 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.

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