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

Nick Sun

April 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 data and resampled real data.

1 Introduction

An important component of maintaining national forests is 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.

Spatial statistics has to deal with what is popularly known as Tobler's First Law: "Everything is related to everything else, but near things are more related than distant things"[5]. Simply put, forestry data at sampled points or plots is likely to be correlated with data points that are close by. 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 the distribution contains a point-mass at value 0 and a positive skewed distribution[6]. 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[4] 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[7], 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.

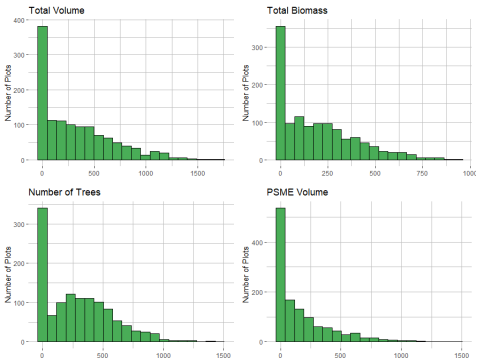
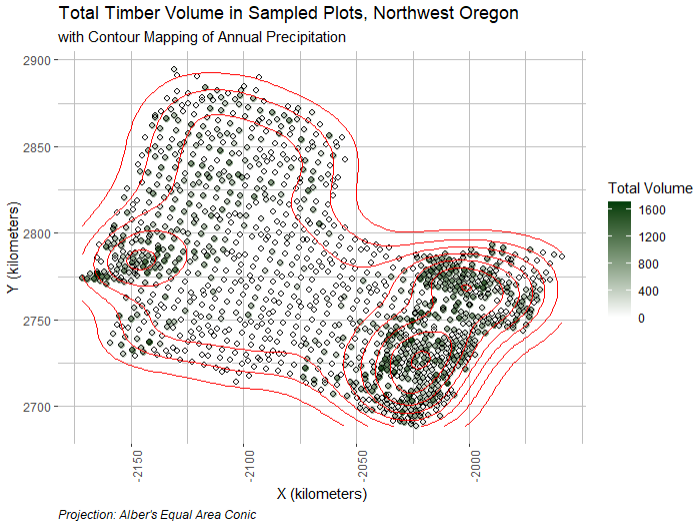


Figure 1: Histograms of response variables.



Simulated Data Using this original forestry dataset, we create $m_1 = 1000$ simulated datasets of size $n = 1224$ 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 hte simulated points.

Resampled Data Again using the original forestry dataset, we obtain $m_2 = 1000$ resampled datasets of size $n = 200$ 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.

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

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.

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

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 obtained by kriging are unbiased and minimize estimation variance.

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 $C(h) = \sigma^2 - \gamma(h)$ where σ^2 is the sample variance of all points.

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, 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[8] and box-office data[3]. 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[4] proposed a spatial Gaussian copula

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

where the correlation matrix Σ is chosen such that it represents the spatial relationships between each of the data points. 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 process for predicting unobserved values using the Gaussian copula model will involve the following steps.

1. Use the observed values to get estimates for the spatial covariance parameters θ_0, θ_1, and the ZIG parameters β, μ_i, and $\pi_i, \forall i$
2. Transform the observed ZIG response variables into standard normal variables using the CDF of the ZIG with the estimated parameters.
3. Use kriging on the standard normal random variables to get estimates for the unobserved values using the formula $\hat{Y}_{unobs} = \Sigma_{obs,unobs}^T \Sigma_{obs}^{-1} Y$
4. Backtransform these unobserved standard normal values to get predictions for the unobserved values on the original scale

3.3 Spatial Random Forest

The random forest is a machine learning algorithm which creates an ensemble of weak learners from randomly selected covariates[2]. 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.

4 Results

Here is my results section where I talk about the outcome of my simulation studies

5 Conclusion

Here is a conclusion where I will probably say something like Gaussian copula is better because statistics over machine learning 5ever.

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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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