A Comparison of the Spatial Linear Model to Nearest Neighbor (k-NN) Methods for Forestry Applications

Jay M. Ver Hoef[1]*, Hailemariam Temesgen[2]

1 Alaska Fisheries Science Center, NOAA Fisheries, Seattle, Washington, United States of America, **2** Department of Forest Engineering, Resources and Management, College of Forestry, Oregon State University, Corvallis, Oregon, United States of America

Abstract

Forest surveys provide critical information for many diverse interests. Data are often collected from samples, and from these samples, maps of resources and estimates of aerial totals or averages are required. In this paper, two approaches for mapping and estimating totals; the spatial linear model (SLM) and k-NN (k-Nearest Neighbor) are compared, theoretically, through simulations, and as applied to real forestry data. While both methods have desirable properties, a review shows that the SLM has prediction optimality properties, and can be quite robust. Simulations of artificial populations and resamplings of real forestry data show that the SLM has smaller empirical root-mean-squared prediction errors (RMSPE) for a wide variety of data types, with generally less bias and better interval coverage than k-NN. These patterns held for both point predictions and for population totals or averages, with the SLM reducing RMSPE from 9% to 67% over some popular k-NN methods, with SLM also more robust to spatially imbalanced sampling. Estimating prediction standard errors remains a problem for k-NN predictors, despite recent attempts using model-based methods. Our conclusions are that the SLM should generally be used rather than k-NN if the goal is accurate mapping or estimation of population totals or averages.

Introduction

Forest surveys provide critical information for many interests: quantifying carbon sequestration, making sound management decisions, designing processing plants, guiding decisions among conflicting land uses, and quantifying wildlife habitats, just to name a few. To meet national and international negotiations and reporting requirements, forest management plans require local inventory data on vegetation, site productivity, biomass, carbon and other resources. The data must be intensive enough to include structural variables relevant to biomass and carbon projections and extensive enough to cover hundreds to thousands of acres, but can not be too expensive to collect. Thus, data are often collected from samples. From these samples, maps of resources and estimates of aerial totals or averages are required. One approach to mapping and estimating totals of biomass and productivity data is a spatial linear model (SLM), which includes ordinary kriging and universal kriging. This approach was initially developed for a similar goal: to predict geographic values or totals for mining resources. However, another approach, k-NN (k-Nearest Neighbor), has been recently developed and has gained widespread use. The overall goal of this paper is to compare SLM to k-NN theoretically, through simulations, and as applied to real forestry data.

The k-NN method finds observed samples that are "close" to an unobserved location based on covariates, and then either imputes the "closest" one directly as a prediction ($k=1$), or forms a weighted average as a prediction ($k>1$). Widespread availability of remotely-sensed data as covariates allows extending ground information to large areas using k-NN. One of the reasons k-NN is popular is that, when $k=1$, predictions are within the bounds of biological reality because they were observed in the samples [1–3]. Also, the logical relationships among response variables will be maintained, so k-NN is a multivariate method that retains the variable relationships seen in the data, particularly when $k=1$ [1,4–6]. When variables are predicted separately, the dependence structure among the response variables is generally lost [7]. The multivariate aspect of the k-NN may be necessary for inventory applications where information on multiple stand attributes is required for stand management decisions or further modeling [5,8]. Because k-NN methods reuse existing samples, they are distribution-free [2,9,10]. Non-parametric k-NN imputation methods may provide better matches to listings of tree species for complex stands with multiple species and a wide variety of tree sizes, which tend to have multi-modal distributions [8]. Non-parametric methods were found to effectively describe local conditions and variability [11,12]. One way to make k-NN local is to select a combination of neighbors from the neighborhood where the average of the covariates is closest to the target record covariates [12,13]. Localization can also be achieved by using spatial coordinates as covariates or by restricting the selection of neighbors to a circular area around the target unit [14]. The yaImpute R package [15] facilitated the comparison of different k-NN approaches and its wide use.

There are some recognized problems with k-NN. It tends to be highly biased at the edge of the data cloud because prediction sites will likely be paired with a more central sample value due to the asymmetric neighborhood [16,17]. Extremely small values and extremely high values will be over- and underestimated, respectively, if the sample data do not cover the whole range of variability [18,19]. Bias can also be a problem in the interior of the data cloud if the covariates are non-uniformly distributed [20]. In k-NN methods, the match is found using the covariates that are available for every site. As the number of samples increases, there will be a higher chance of getting an exact match in the covariates. However, this only guarantees an exact match for the response variables if there is perfect (or very strong) correlations with the covariates. As sample size increases, there is no guarantee that the mean of the response variables will approach the true mean and hence k-NN methods are not statistically consistent [3]. The k-NN methods lack a good measure of uncertainty, and often the global root-mean squared error from cross-validation is used for point-wise standard errors [20]. However, that global root-mean squared error may not be a good measure of accuracy when response variables have heteroscedastic variances around the covariates; it was recommended that graphical tools be used to evaluate issues of bias, homoscedasticity, influential observations, outliers, and extrapolations [21]. An approach that uses a model of the covariates space variogram to quantify prediction uncertainty was proposed, but is computationally demanding [22]. Relevant accuracy statistics for assessing the quality of predictions of categorical variables are still lacking [7]. However, there are recent proposals for a variance estimator that incorporates spatial correlation [23], model-based estimators of the uncertainty [24], and design-based approaches to derive the statistical properties of the k-NN predictions [17]. If $k > 1$, then k-NN loses many of its purported benefits as some estimates may not be within the realm of real values [3]. If there are several response variables or "rare" polygons (stands), a good match will be very difficult to find [3].

Geostatistical methods were also devised for prediction, both for single sites or block averages. For a history, see [25]. Combining the notion of classical geostatistics with a linear model (e.g., regression) yields the spatial linear model (SLM) (called universal kriging in the geostatistical literature). In comparison to k-NN, SLM predictions were designed to minimize the root-mean-squared error. If the data-generating process is multivariate normal, then the SLM predictor is equivalent to the conditional expectation, which is optimal [26](pgs. 108–110). Moreover, if the generating mechanism is not multivariate normal, the SLM predictions are still optimal among the class of linear predictors, called best linear unbiased prediction (BLUP) [26](pgs. 151–155). The BLUP was extended to finite populations of spatial data [27–29]. Although distribution-free, BLUP requires specification of a covariance model; however predictions are generally robust to mis-specification of the covariance model [30,31]. Classically, geostatistical methods estimated the covariance model by binning data into distance classes and using a least-squares fit [32]; however, restricted maximum likelihood (REML) estimation [33,34] removes the arbitrary nature of binning and fitting and is an unbiased estimating equation approach to the SLM [35,36]. Finally, SLM predictors are linear, creating weighted averages of data, and we can appeal to a spatially correlated version of the central limit theorem that predictors will be asymptotically normally distributed, e.g. [37], allowing for inference based on a standard normal distribution (e.g., for prediction intervals). More detailed and mathematically-oriented arguments on the stability of the geostatistical method can be found in [38] and [26](pgs. 289–

299). There are many extensions of the SLM when predictions cannot be assumed approximately normally-distributed, e.g. [39].

The SLM is not without some recognized problems. Most of these are related to the problem of estimating the spatial covariance function. In a sense, the data are used twice; once to estimate the covariance parameters and secondly used for BLUP. This has been termed empirical best linear unbiased prediction (EBLUP) [40]. Forming multivariable models requires estimating many covariance parameters, and any gain in precision is often minimal [41], with gains often 10% or less [42]. However, see [43,44] for forestry applications. Some claims against the SLM are in error. Some forestry data are obtained from polygons that do not have a unique position in space and are irregular lattice data (aggregates of spatial data). Some authors, e.g. [6], indicate that the SLM is not appropriate for these data, but that is not correct. Polygon centroids can be used to compute distance, as shown with the real forestry data below. We note, however, that this may be a problem for very irregularly shaped polygons.

It turns out that predictors for both SLM and k-NN are linear, and the above review indicates that SLM should be optimal. However, much of that theory is based on an assumed spatial stochastic model, does not take into account the estimation of the covariance parameters, nor does it say how much better SLM might be. k-NN has fewer assumptions, and perhaps it is easier to implement and faster to compute, and a small loss in efficiency is compensated by ease of use and computational speed (see [45] and references therein). We build on a few previous comparisons [46,47], where they compare only for point predictions, they do not assess the validity of prediction standard errors, and they use variogram estimators on residuals, which is known to be biased for covariance parameters [48](pgs. 257–258). One problem in making the comparison is that k-NN is an algorithm that makes no assumption about how spatial data were created; it only assumes a fixed surface. The SLM, on the other hand, is based on the idea that data are a realization of a spatial stochastic model. By conditioning on realizations under a SLM, and hence adopting a fixed surface perspective, k-NN and SLM can be evaluated in a common framework. Foresters and forest managers are interested in the global performance of predictive or imputation methods across a management region, among management regions, and across time. Supporting that practice, the comparison of prediction or imputation methods using global computations of average bias and RMSPE (described in Section "Performance Measures") is deeply rooted in the forestry literature [3,10,14,21].

Our goal is to compare k-NN to the SLM using simulations of artificial populations and resampling real forestry data. In the Methods section, we make explicit the prediction goals and k-NN and SLM methods and models, and we describe the simulation methods and real forestry data. The outcomes of using k-NN and SLM on these data are presented in the Results section. We discuss the results and offer some conclusions in the Discussion section.

Methods

We will use the following notation for spatial data. Let the population of response values be partitioned into those that are observed $\mathbf{y}_o = \{y_i; i = 1,...,n\}$ and those that are unobserved $\mathbf{y}_u = \{y_i; i = n+1,...,n+m\}$, and $\mathbf{y} = (\mathbf{y}'_o, \mathbf{y}'_u)'$. Let the index set for the observed data be $\mathbb{O} = 1, . . . ,n$ and for the unobserved data be $\mathbb{U} = n+1, . . . ,n+m$. We consider two main goals: 1) point prediction of y_j for $j \in \mathbb{U}$, and 2) block prediction of the total or average $T = \sum_{i=1}^{n+m} b_i y_i = \mathbf{b}' \mathbf{y}$, where b_i are the weights that define the block objective; e.g., if $\{b_i = 1; i = 1, . . . ,(n+m)\}$ then T is a population total, and if $\{b_i = 1/(n+m); i = 1, . . . ,(n+m)\}$ then T

is a population average. Note that prediction goals for small areas can also be defined using zeros as weights in $\mathbf{b}$, but we do not pursue that here. For all response values, there are covariates contained in a design matrix $\mathbf{X}$ which has $n+m$ rows and p columns, where the first n rows, $\mathbf{X}_o$, correspond to $\mathbf{y}_o$ and the next m rows, $\mathbf{X}_u$, correspond to $\mathbf{y}_u$. The spatial coordinates are contained in the matrix $\mathbf{S}$ which has $n+m$ rows and q columns, where the first n rows correspond to $\mathbf{y}_o$ and the next m rows correspond to $\mathbf{y}_u$. In $\mathbf{S}$, q is generally two. For example, the first column could be longitude and the second column latitude, or some planar transformation of them. A coordinate such as height might also be included, but we do not consider it here.

To meet our two objective above, we define the linear predictor,

$$\hat{y}_j = \sum_{i\in\mathbb{O}} \lambda_{i,j} y_i = \lambda_j' \mathbf{y}_o, \tag{1}$$

where $j\in\mathbb{U}$. A linear block predictor is

$$\hat{T} = \sum_{i\in\mathbb{O}} b_i y_i + \sum_{j\in\mathbb{U}} b_j \hat{y}_j = \omega' \mathbf{y}_o. \tag{2}$$

As will be seen, both k-NN and SLM are linear predictors. Note that [23] and [24] attempt model-based estimators for a version (2) that includes predicted values of $y_i; i\in\mathbb{O}$ rather than simply summing the observed values. The variance of such predictors lack any notion of a finite population that shrinks the variance as the sampling fraction increases; i.e., we want a variance estimator such that $var(\hat{T})=0$ if the whole population is observed, so we do not pursue their formulation any further.

Both the SLM and k-NN use distance in various ways so a general definition is given here. Let $\mathbf{A}$ be a matrix with coordinates in the columns and the ith row denoted as $\mathbf{a}_i$. A general distance formula between the ith and jth rows of $\mathbf{A}$ is,

$$d(i,j;\mathbf{A},\mathbf{W}) \equiv \sqrt{(\mathbf{a}_i - \mathbf{a}_j)' \mathbf{W}(\mathbf{a}_i - \mathbf{a}_j)}, \tag{3}$$

where $\mathbf{W}$ is a weighting matrix.

Review of k-NN

Let $\mathbf{x}_i'$ and $\mathbf{x}_j'$ be the ith and jth rows of $\mathbf{X}$, respectively. Then a "variable-space distance" between the ith and jth sites can be computed as $d(i,j;\mathbf{X},\mathbf{W})$. Several types of distances are possible. For example, if $\mathbf{W}$ is the identity matrix, this is raw distance; if $\mathbf{W}$ is diagonal with the inverse of the empirical variance for each of the columns in $\mathbf{X}$ as the diagonal elements, then this is normalized distance; if $\mathbf{W}$ is the inverse of the empirical covariance matrix among the columns in $\mathbf{X}$, then this is Mahalanobis distance; and if $\mathbf{W} = \Gamma \Lambda^2 \Gamma'$, where Γ is the matrix of canonical vectors from canonical correlation analysis between $\mathbf{X}$ and $\mathbf{y}$ and Λ is the canonical correlation matrix, then this is the "most similar neighbor" (MSN) distance [1]. All of these and others have been implemented in the yaImpute package [15] in R [49]. The k-NN method chooses weights λ_j based on a distance matrix. Let $\mathbf{D}$ be a distance matrix with i,jth element $d(i,j;\mathbf{X},\mathbf{W})$, which can be partitioned as

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_{o,o} & \mathbf{D}_{o,u} \\ \mathbf{D}_{o,u}' & \mathbf{D}_{u,u} \end{bmatrix}$$

Let $\mathbf{d}_j$ be the jth column of $\mathbf{D}$, $j\in\mathbb{U}$, contained in $\mathbf{D}_{o,u}$; i.e., $\mathbf{d}_j = \{\mathbf{D}[i,j]; i=1,2,\ldots,n; j\in\mathbb{U}\}$. If i is the index for $\min(\mathbf{d}_j)$, then for a first-order nearest neighbor, $\lambda_{i,j}=1$ in (1) and all other $\lambda_{\ell,j}=0; \ell\neq i$. This essentially assigns the value of y_i to $\hat{y}_j$ for the ith site that is closest to the jth site in variable-space. Let $\mathbb{N}_j^k$ be the index set of the k nearest sites (smallest values) in $\mathbf{d}_j$. Then $\lambda_{i,j}=1/k; i\in\mathbb{N}_j^k$ takes the average of $\{y_i\}$ from the k nearest neighbors in variable-space. Another option is to weight inversely proportional to distance, where $\lambda_{i,j}\propto 1/(1+d_{i,j}); i\in\mathbb{N}_j^k$ where $\sum_{\mathbb{N}_j^k} \lambda_{i,j}=1$ and $d_{i,j}$ is the ith element in $\mathbf{d}_j$.

Cross-validation is the method most often used to compute prediction standard errors [20]. Cross-validation makes predictions for sites that already have values, where each sample is removed one at a time, and the rest of the sample is used to predict the one that was removed. The idea is to use in-sample averaged squared errors between the predicted and observed values to serve as a global estimator of squared errors when out-of-sample. Let the k-NN prediction standard error be estimated as,

$$\hat{\xi} = \sqrt{\frac{1}{n}\sum_{i=1}^n (\hat{y}_i - y_i)^2}, \tag{4}$$

where $\hat{y}_i$ is the cross-validation prediction of y_i for $i=1,\ldots,n$ sample values. Assuming prediction errors are normally-distributed, 90% prediction intervals are formed as $\hat{y}_j \pm 1.645\hat{\xi}$ for $j=n+1,\ldots,n+m$ out-of-sample values. Note that these are constant for all j; we examine a spatially explicit model-based approach [23] in Section "A geostatistical approach for estimating the variance of k-NN predictors."

For the standard error of estimating a total, we borrow from the idea of classical sampling theory, e.g. [50], where $\hat{\hat{\xi}}$ replaces the standard error,

$$se(\hat{T}) = \hat{\hat{\xi}}\sqrt{(n+m)m}. \tag{5}$$

Review of SLM

Assume only the linear model

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \tag{6}$$

where $\mathbf{X}$ is a matrix of fixed covariates, β is a vector of parameters, and ϵ is a random vector with $var(\epsilon)=\mathbf{V}$ for some unknown spatial multivariate distribution. Note the contrast to k-NN; (6) is a spatial stochastic model that allows optimization with respect to bias and squared error, which we now review. Let $\mathbf{V}$ be partitioned as

$$\text{var}(\mathbf{y}) = \mathbf{V} = \begin{bmatrix} \mathbf{V}_{o,o} & \mathbf{V}_{o,u} \\ \mathbf{V}_{o,u}' & \mathbf{V}_{u,u} \end{bmatrix}, \tag{7}$$

and let $\mathbf{v}_j$ be the jth column of $\mathbf{V}$, $j\in\mathbb{U}$, contained in $\mathbf{V}_{o,u}$; i.e., $\mathbf{v}_j = \{\mathbf{V}[i,j]; i=1,2,\ldots,n; j\in\mathbb{U}\}$. Consider squared-error loss, and the linear predictor λ_j in (1), where $\lambda_j'\mathbf{X}_o = \mathbf{x}_j'$ and $\mathbf{X}_0$ contains at least a column of ones. The predictor that minimizes squared-error loss, known as the best linear unbiased predictor (BLUP), $\hat{y}_j = \lambda_j'\mathbf{y}_o$, has

$$\lambda_j' = [\mathbf{v}_j + \mathbf{X}_o\mathbf{C}(\mathbf{x}_j - \mathbf{X}_o'\mathbf{V}_{o,o}^{-1}\mathbf{v}_j)]'\mathbf{V}_{o,o}^{-1}, \qquad (8)$$

where $\mathbf{C} = (\mathbf{X}_o'\mathbf{V}_{o,o}^{-1}\mathbf{X}_o)^{-1}$, with prediction variance of

$$\mathrm{var}(\hat{Y}_j - Y_j) = \mathbf{V}[j,j] - 2\lambda_j'\mathbf{v}_j + \lambda_j'\mathbf{V}_{o,o}\lambda_j \qquad (9)$$

[26] (pgs. 151–155). Notice that β is unknown; the only assumption is the linear model and a known spatial covariance matrix.

Assume the same linear model in (6), except this time the linear predictor is (2). Let $T = \mathbf{b}'\mathbf{y}$, where $\mathbf{b}$ is a vector of all ones, and $\mathbf{b}_o = \{b_i; i=1:n\}$ and $\mathbf{b}_u = \{b_i; i=(n+1):(n+m)\}$. For a finite population the BLUP, $\hat{T} = \omega'\mathbf{y}_o$, has

$$\omega' = \mathbf{b}_o' + \mathbf{b}_u'[\mathbf{V}_{o,u}' - \mathbf{V}_{o,u}'\mathbf{V}_{o,o}^{-1}\mathbf{X}_o\mathbf{C}\mathbf{X}_o' + \mathbf{X}_u\mathbf{C}\mathbf{X}_o']\mathbf{V}_{o,o}^{-1}, \quad (10)$$

with prediction variance of

$$\mathrm{var}(\hat{T} - T) = \mathbf{b}_u'[\mathbf{V}_{u,u} - \mathbf{V}_{o,u}'\mathbf{V}_{o,o}^{-1}\mathbf{V}_{o,u} + \mathbf{F}'\mathbf{C}\mathbf{F}]\mathbf{b}_u, \qquad (11)$$

where $\mathbf{F} = \mathbf{X}_u' - \mathbf{X}_o'\mathbf{V}_{o,o}^{-1}\mathbf{V}_{o,u}$ [27–29]. The finite population correction factor is not obvious in (11). However, as $\mathbf{b}_u$ gets shorter in length, (11) goes to zero. If $\mathbf{V} = \sigma^2\mathbf{I}$, then (11) simplifies to $\mathrm{var}(\hat{T}) = (n+m)^2\sigma^2(1-f)/n$ where $f = n/(n+m)$ is the sampling fraction; this is the classical formula in simple random sampling without replacement for finite populations, e.g. [50](pg. 16).

For equations (9) and (11) $\mathbf{V}$ is unknown and must be estimated. In spatial models, $\mathbf{V}$ is modeled through spatial information; in geostatistics this is spatial distance. Consider the exponential autocovariance model,

$$\mathbf{V}[i,j; \theta] = cov(Y_i, Y_j) = \sigma^2 + \delta^2 G(d_{i,j}; \rho) \qquad (12)$$

where $G(d_{i,j}: \rho)$ is a general autocorrelation function, $d_{i,j} = d(i,j;\mathbf{S},\mathbf{I})$ as defined in (3), $\mathbf{V}[i,j; \theta]$ is the i,jth element of $\mathbf{V}$, $\theta = (\delta^2, \rho, \sigma^2)$ with δ^2 as the partial sill, ρ as the range parameter, and σ^2 as the nugget effect (which may absorb spatial autocorrelation at very fine scales within minimum sampling distances). We will fit models using $G(d_{i,j}) = \exp(-d_{i,j}/\rho)$; for many other models see [51](pgs. 80–96). The larger ρ, the more autocorrelation between sites for a given distance. The parameters σ^2 and δ^2 are variance components, with δ^2 controlling the autocorrelated component and σ^2 controlling the uncorrelated component. For all models in this article, we use (12), relying on the fact that inferences are generally robust to mis-specification of the model. We estimate the covariance parameters using restricted maximum likelihood (REML) [33,34],

$$l(\theta; \mathbf{y}_o) = log|\mathbf{V}_{o,o}(\sigma^2,\rho)| + \mathbf{r}'\mathbf{V}_{o,o}^{-1}(\sigma^2,\rho)\mathbf{r} \\ + log|\mathbf{X}_o'\mathbf{V}_{o,o}^{-1}(\sigma^2,\rho)\mathbf{X}_o| + c, \qquad (13)$$

where $\mathbf{r} = \mathbf{y}_o - \mathbf{X}_o\hat{\beta}$, $\hat{\beta} = (\mathbf{X}_o'\mathbf{V}_{o,o}^{-1}(\theta)\mathbf{X}_o)^{-1}\mathbf{X}_o'\mathbf{V}_{o,o}^{-1}(\theta)\mathbf{y}_o$, the dependence of $\mathbf{V}_{o,o}$ on θ is denoted as $\mathbf{V}_{o,o}(\theta)$, and c is a constant that does not depend on θ. Equation (13) is an unbiased estimating equation [35,36] and minimizing it for σ^2 and ρ provides their REML estimates. Using the estimated covariance parameters from REML in equations (8–11) provides the EBLUP predictors and standard errors.

The significance of using (13) is three-fold: 1) normality is not required to use (13) because it is an unbiased estimating equation, 2) there is no need to de-trend because estimation of β is essentially embedded in (13), and 3) there is no need to compute empirical variograms by binning residuals. Residuals from de-trending are biased [48](pgs. 257–258) and binning is arbitrary. Thus, (13) provides an automatic way to estimate a spatial covariance matrix in very general conditions.

A Geostatistical Approach for Estimating the Variance of k-NN Predictors

An iterated variogram estimator for the variance of a k-NN predictor has been proposed [23]. Suppose that we start with iteration $q = 0$ and (4) as an estimator for a constant prediction standard error, $\hat{v}_i^{[q]} = \hat{\xi}$. Form standardized residuals as

$$u_i^{[q]} = (y_i - \hat{y}_i)/\hat{v}_i^{[q]}, \qquad (14)$$

where $\hat{y}_i$ is the in-sample cross-validation prediction value using some k-NN method. Then compute an empirical semivariogram,

$$\hat{\gamma}^{[q]}(h_\ell) = \frac{\sum_i^n \sum_{i'}^n I(d_{i,i'}\in N_\ell)(u_i^{[q]} - u_j^{[q]})^2}{2\sum_i^n \sum_{i'}^n I(d_{i,i'}\in N_\ell)} \qquad (15)$$

where $d_{i,i'} = d(i,i';\mathbf{S},\mathbf{I})$, the ℓth distance class is $N_\ell = (t_{\ell-1}, t_\ell]$, where $t_0 = 0$, $0 < t_1 < t_2 <,\ldots,< t_L$, and h_ℓ is some function of $\mathbb{D}_\ell = \{d_{i,i'}; d_{i,i'}\in N_\ell\}$; e.g., h_ℓ might be mean of all distances in $\mathbb{D}_\ell$, the median of $\mathbb{D}_\ell$, or the midpoint $h_\ell = (t_\ell - t_{\ell-1})/2$. A semivariogram, such as the equivalent to (12),

$$\gamma(h_\ell; \theta) = \sigma^2 + \delta^2(1 - \exp(-d_{i,i'}/\rho)), \qquad (16)$$

is fit to (15), typically by minimizing a weighted-least-squares criteria; e.g. [32],

$$\hat{\theta}^{[q]} = \underset{\theta}{\mathrm{argmin}} \sum_{\ell=1}^{L} |\mathbb{D}_\ell| \left(\frac{\hat{\gamma}^{[q]}(h_\ell)}{\gamma(h_\ell; \theta)} - 1\right)^2. \qquad (17)$$

Let the spatial autocorrelation be $c_{i,i'}^{[q]} = \exp(-d_{i,i'}/\rho^{[q]})$. Then a local estimator of variance is

$$v_j^{[q+1]} = \frac{\sum_{i\in\mathbb{N}_j^k}(y_i-\hat{y}_j)^2}{k-(1/k)\sum_{i\in\mathbb{N}_j^k}\sum_{i'\in\mathbb{N}_j^k}c_{i,i'}^{[q]}}.$$

Now go back to (14) with updated v's and iterate until convergence. For a convergence criteria, we used

$$\max_{j\in\mathbb{U}}\{(v_j^{[q+1]}-v_j^{[q]})/v_j^{[q]})\}<10^{-5}.$$

After convergence, a local estimator of prediction variance is

$$\text{var}(\hat{y}_j-y_j) = \frac{v_j^2}{k^2}\left(\sum_{i\in\mathbb{N}_j^k}\sum_{i'\in\mathbb{N}_j^k}c_{i,i'}-2k\sum_{i\in\mathbb{N}_j^k}c_{i,j}+k^2\right), \quad (18)$$

where the iteration superscript $[q]$ is suppressed after convergence. The prediction variance estimator (18) can be seen as an attempt to form a local (in covariate space) version of (9) without having to estimate mean effects due to "nearness" in covariate space. Note that this estimator will not work for $k=1$, and it is only sensible when the mean of nearest neighbors is computed, as compared to the distance-weighted version. However, the above method could be adopted for distance-weighted. The estimator (18) will be examined using simulations. For the simulations, we used 10 equal-interval variogram distance-bins ($t_\ell-t_{\ell-1}$ was equal for all ℓ; $L=10$) between 0 and the maximum distance in the data set. Equation (17) was minimized using the optim() function in R [49] with the Nelder-Mead simplex method [52], obtaining a starting value from a $10\times10\times10$ search grid for the parameters of (16). A maximum of 30 iterations was allowed.

A more formal geostatistical analysis of the k-NN predictors is as follows. The k-NN predictor can be written as (1). Under the SLM, the root-mean-squared prediction error (RMSPE) is

$$E(\hat{Y}_j-Y_j)^2 = E[(\boldsymbol{\lambda}_j'(\mathbf{X}_o\boldsymbol{\beta}+\boldsymbol{\epsilon}_o)-(\mathbf{x}_j'\boldsymbol{\beta}+\epsilon_j))^2].$$

After taking expectations,

$$E(\hat{Y}_j-Y_j)^2 = [(\boldsymbol{\lambda}_j'\mathbf{X}_o-\mathbf{x}_j')\boldsymbol{\beta}]^2 + \boldsymbol{\lambda}_j'\mathbf{V}_{o,o}\boldsymbol{\lambda}_j + \boldsymbol{\lambda}_j'\mathbf{v}_j+\mathbf{V}[j,j].$$

Note that k-NN methods make the sensible constraint that $\boldsymbol{\lambda}_j'\mathbf{1}=1$ when using the mean of the nearest neighbors or distance-weighting. If $\mathbf{X}_o$ is a single column of ones, then the bias term $[(\boldsymbol{\lambda}_j'\mathbf{X}_o-\mathbf{x}_j')\boldsymbol{\beta}]^2$ above disappears. However, this is not true in general. In contrast, under the BLUP, the bias term disappears due to further constraints on $\boldsymbol{\lambda}_j$ that guarantee unbiasedness for any $\mathbf{X}_o$ and $\mathbf{x}_j'$. Hence, the RMSPE for k-NN will be greater than BLUP for two reasons: it is not optimized for minimizing the error variance $\boldsymbol{\lambda}_j'\mathbf{V}_{o,o}\boldsymbol{\lambda}_j + \boldsymbol{\lambda}_j'\mathbf{v}_j+\mathbf{V}[j,j]$ and there is a bias-squared component.

An estimate of the RMSPE of the k-NN predictor under the SLM model is obtained by replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$. Note that

$$E[(\boldsymbol{\lambda}_j'\mathbf{X}_o-\mathbf{x}_j')\hat{\boldsymbol{\beta}}-(\boldsymbol{\lambda}_j'\mathbf{X}_o-\mathbf{x}_j')\boldsymbol{\beta}+(\boldsymbol{\lambda}_j'\mathbf{X}_o-\mathbf{x}_j')\boldsymbol{\beta}]^2,$$

which equals

$$E[(\boldsymbol{\lambda}_j'\mathbf{X}_o-\mathbf{x}_j')(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\boldsymbol{\lambda}_j'\mathbf{X}_o-\mathbf{x}_j')'] + [(\boldsymbol{\lambda}_j'\mathbf{X}_o-\mathbf{x}_j')\boldsymbol{\beta}]^2$$

because $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$. Note that $E(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})' = (\mathbf{X}_o'\mathbf{V}_{o,o}^{-1}\mathbf{X}_o)^{-1} = \mathbf{C}$, so

$$\sqrt{[(\boldsymbol{\lambda}_j'\mathbf{X}_o-\mathbf{x}_j')\hat{\boldsymbol{\beta}}]^2 + \boldsymbol{\lambda}_j'\mathbf{V}_{o,o}\boldsymbol{\lambda}_j + \boldsymbol{\lambda}_j'\mathbf{v}_j+\mathbf{V}[j,j]-(\boldsymbol{\lambda}_j'\mathbf{X}_o-\mathbf{x}_j')\mathbf{C}(\boldsymbol{\lambda}_j'\mathbf{X}_o-\mathbf{x}_j')'}$$

(19)

is an estimator of the RMSPE of a k-NN predictor using parameters estimated under a SLM; i.e., covariance parameters can be estimated using (13) and $\boldsymbol{\beta}$ can be estimated using generalized least squares, $\hat{\boldsymbol{\beta}}=\mathbf{C}\mathbf{X}_o'\mathbf{V}_{o,o}^{-1}\mathbf{y}_o$. We will analyze some k-NN estimators using (19) in simulations below.

Simulation of Artificial Data

We created spatially-patterned and cross-correlated X-variables. All data sets were repeatedly simulated on a 20×20 regular grid evenly spaced between -1 and 1 on both coordinate axes and eight covariates: X_1-X_8, described next. Start with

$$\mathbf{w}_1 = \mathbf{z}_1 + \boldsymbol{\epsilon}_1,$$

where $\mathbf{z}_1$ is a 400×1 vector of values (on the 20×20 regular grid) containing zero-mean spatially-autocorrelated random variables from some geostatistical model with partial sill δ_1^2 and range parameter ρ_1, and $\boldsymbol{\epsilon}_1$ is a 400×1 vector containing zero-mean independent random variables with variance σ_1^2. We also let $\mathbf{z}_1$ be independent of $\boldsymbol{\epsilon}_1$. Next, we set up an autoregressive-like recursion, where

$$\mathbf{w}_\eta = \phi_{\eta-1}\mathbf{w}_{\eta-1}+\mathbf{z}_\eta + \boldsymbol{\epsilon}_\eta,$$

where $\mathbf{z}_\eta,\eta=2,3,\ldots,8$ contains zero-mean spatially-autocorrelated random variables from some geostatistical model with partial sill δ_η^2 and range parameter ρ_η, and $\boldsymbol{\epsilon}_\eta$ contains zero-mean independent random variables with variance σ_η^2, where again $\mathbf{z}_\eta$ is independent of $\boldsymbol{\epsilon}_\eta$. Note that $\phi_{\eta-1}$ is a parameter that creates cross-correlation between variables by regressing $\mathbf{w}_\eta$ on $\mathbf{w}_{\eta-1}$. This set-up ensures cross-correlation among the W-variables and spatial autocorrelation within each W-variable. Now let

$$\mathbf{x}_\eta = \boldsymbol{\mu}_\eta+\mathbf{w}_\eta,$$

where all the elements of $\boldsymbol{\mu}_\eta$ are constant, equal to μ_η. Finally, create the response variable as,

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{z}_y + \boldsymbol{\epsilon}_y, \quad (20)$$

where $\mathbf{X}=[\mathbf{x}_1|\mathbf{x}_2|\ldots|\mathbf{x}_8]$, $\boldsymbol{\beta}$ is a vector of parameters, and $\mathbf{z}_y$ contains zero-mean spatially-autocorrelated random variables

from some geostatistical model with partial sill δ_y^2 and range parameter ρ_y, and ϵ_y contains zero-mean independent random variables with variance σ_y^2, where again $\mathbf{z}_y$ is independent of ϵ_y.

For simulations, we let all $\mathbf{z}_\eta$ and ϵ_η be normally distributed. For autocorrelation of $\mathbf{z}_\eta$, we used the spherical model,

$$\mathbf{V}[j,j^{'};\sigma^2,\rho] = \sigma^2\left(1 - \frac{3}{2}\frac{d_{j,j^{'}}}{\rho} + \frac{1}{2}\frac{d_{j,j^{'}}}{\rho^3}\right)I\left(\frac{d_{j,j^{'}}}{\rho} \le 1\right), \quad (21)$$

where $d_{j,j^{'}} = d(j,j^{'};\mathbf{S},\mathbf{I})$

We simulated three types of data using these models. In all cases $\phi_\eta = (NA, 0.5, 0.5, 0.5, 0, 0.5, 0.5, 0.5)$. Note that for each simulated data set the covariates $X_1 - X_4$ were cross-correlated through $\phi_2 - \phi_4$, but $\phi_5 = 0$ broke any further cross-correlation to the group $X_5 - X_8$, and then $X_5 - X_8$ as a group were cross-correlated.

1. The first simulated method had $\rho_\eta^2 \in \{0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2\}$, $\delta_\eta^2 \in \{1,2,\ldots,8\}$, $\mu_k \in \{1,2,\ldots,8\}$ and $\sigma_\eta^2 \in \{0.1,\ldots,0.1\}$, $\beta = (-3,-2,\ldots,2,3)$, $\rho_y = 3$, $\delta_y^2 = 20$, and $\sigma_y^2 = 1$.

2. For the second simulation method, let $\mathbf{y}_t$ denote the simulated data, with $\rho_\eta^2 \in \{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4\}$, $\delta_\eta^2 \in \{1/8, 2/8,\ldots,1\}$, $\mu_\eta \in \{-.4,-.2,0,.2,.4,.6,.8,1\}$ and $\sigma_\eta^2 \in \{0.01,\ldots,0.01\}$, $\beta = (-3/6,-2/6,\ldots,2/6,3/6)$, $\rho_y = 4$, $\delta_y^2 = 1$, and $\sigma_y^2 = .05$. Then, $\mathbf{y}$ was simulated from $\text{Poi}(\exp(\mathbf{y}_t))$ where $\text{Poi}(\cdot)$ is a Poisson distribution. Note that $\mathbf{y}_t$ wandered well below zero over significant spatial patches creating zero-inflated count data. Variances and regression coefficients are smaller here to keep numbers reasonable when exponentiating.

3. For the third simulation method, let $\mathbf{y}_t$ denote the simulated data, with all of the same parameters as for the second simulation. Then, $\mathbf{y}$ was simulated as $y_j = I(y_{t,j} > 0)$, creating binary data of zeros and ones.

Note that the group $X_1 - X_4$ had smaller ranges and variances than group $X_5 - X_8$. The response variable was related to both groups through β, but the coefficients were zero for X_4 and X_5. We made predictions for both k-NN and SLM methods using only X_1, X_2, X_4, X_5, X_7, X_8. Thus, several model mis-specifications were made: inclusion of covariates with no effect (X_4 and X_5), exclusion of covariates with real effects (X_3 and X_6), and in the case of SLM, generating data from a spherical autocovariance model but fitting the data with an exponential autocovariance model.

For each of the three simulation methods listed above, 2000 data sets were simulated (with 400 simulated values per data set). For each simulation 100 locations were sampled randomly, and from each sample the other 300 were predicted, along with the overall total.

Forest Data

Data for this study were obtained from the Forest Inventory and Analysis (FIA) databases for Oregon. The FIA databases are part of the national inventory of forests for the United States [53,54]. The Food Security Act of 1985 protects the confidentiality of the Forest Inventory and Analysis (FIA) plot locations, the integrity of the FIA sample, and the privacy of landowners who allow FIA field crews on their land. For these reasons, the actual plot locations are not available publicly. Persons needing exact plot

locations to examine or reproduce our results should contact the Pacific Northwest Forest Inventory and Analysis Program at http://www.fs.fed.us/pnw/fia/. Alternatively, FIA produces and maintains a set of public databases with perturbed plot coordinates that can be downloaded and used by anyone. The website for accessing these data is http://www.fia.fs.fed.us/tools-data/.

A tessellation of hexagons, each approximately 2400 hectares in size, is superimposed across the nation, with one field plot randomly located within each hexagon. Approximately the same number of plots is measured each year, each plot has the same probability of selection, and in the western U.S. plots are remeasured every ten years. Each field plot is composed of four subplots. Forested areas that are distinguished by structure, management history, or forest type are mapped as unique polygons (also called condition-classes) on the plot and correspond to stands of at least 0.4047 hectare in size. For our study area there were 1886 forested FIA plots measured between 2001 and 2006. Biomass (DRYBIOT), maximum potential mean annual increment (PMAI), elevation, and primary species identifier (i.e., Douglas-fir (*Pseudotsuga menziesii*), ponderosa pine (*Pinus ponderosa*) or western hemlock (*Tsuga heterophylla*)) were obtained from the FIA annual database.

For climate data, we used monthly temperature and precipitation normal data for the period 1971–2000 produced by the Parameter-elevation Regressions on Independent Slopes Model (PRISM), which provided an 800 m grid that produced differences between measured plot elevation and overlaid PRISM grid elevation up to 350 mm in the mountainous areas of Oregon. To account for changes in climate due to these elevation differences we utilized a process similar to [55] where we created a scale-free interpolation process using a 90 m digital elevation model and PRISM temperature and elevation gradients of the larger 800 m grid. The result is a 90 m monthly climate grid. Like [55] we used this procedure for extracting temperature (T) only and used a simple distance weighting method for precipitation (P). Climate Moisture Index (CMI), a measure of precipitation in excess of evapotranspiration, was used to quantify moisture availability. The raw values of PMAI and DRYBIOT, along with residuals after fitting a linear model, are shown in Figure 1; covariates (X-variables) in the linear model were temperature, precipitation, Climate Moisture Index (CMI), WH (an indicator variable for shade tolerance based on Western Hemlock trees) and elevation.

For real forest data, we used a census of 1886 known FIA values for PMAI and DRYBIOT. These were subsampled in three different experiments. For each resampling experiment listed below, the data sets were subsampled 500 times. For each subsample, a sample of 386 values was chosen randomly without replacement, and from each sample the remaining 1500 locations were predicted, along with an estimate of the total for all 1886 values. The covariates used were temperature, precipitation, CMI, WH, and elevation. The 500 resamplings were performed for each of the following:

1. PMAI: this variable reflects the maximum potential forest productivity at a particular site. It indicates the average annual productivity of wood volume (m^3/ha/year) that would be realized over time.

2. DRYBIOT: this variable is the total above-ground oven-dry biomass of live trees larger than 2.5 cm in diameter.

3. PMAI with unbalanced spatial sampling. Here, we preferentially sampled geographically, as often happens in real applications due to access or other issues. We divided up the study area into four parts, using $x = 100$ and $y = 100$ as the
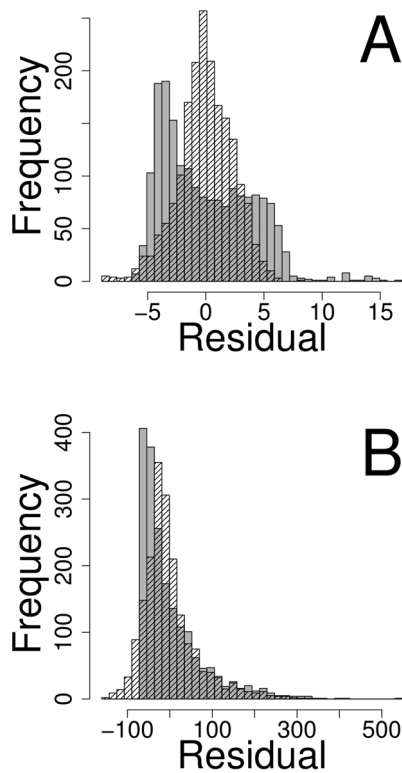
Figure 1. Histograms of A. PMAI, B. DRYBIOT. The gray-shaded histograms are based on the original centered data, and the cross-hatched histogram is based on the residuals after fitting a multiple regression model with main effects for all covariates.
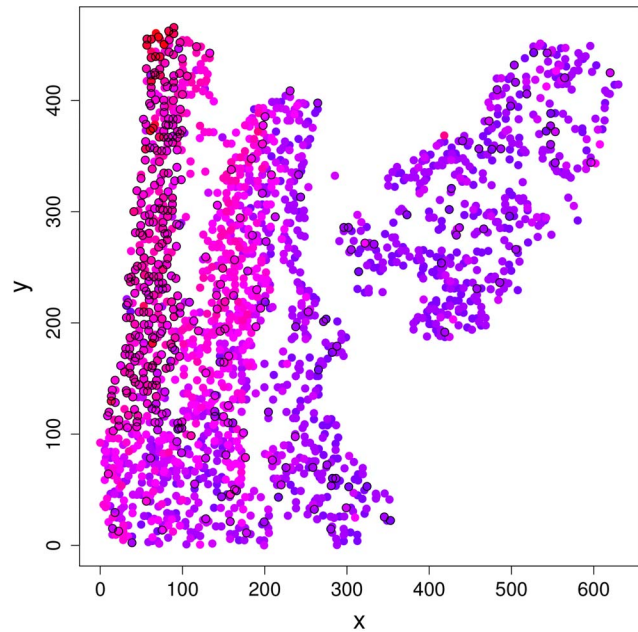doi:10.1371/journal.pone.0059129.g001

Figure 2. Spatial locations of PMAI variable. The redder shades indicate higher values, and the bluer shades indicate lower values. One draw from the unbalanced spatial sample is shown with black circles around the sampled locations.
doi:10.1371/journal.pone.0059129.g002

dividing lines. We created samples by sampling 200 values without replacement for $y > 100$ and $x < 100$, and then randomly sampling 186 values without replacement from the rest of area. One such sample is shown in Figure 2. From each unbalanced spatial sample, the remaining 1500 locations were predicted.

Performance Measures

Over the six preceding simulation/resampling experiments, the SLM is compared to k-NN using three measures of predictive performance, both for individual site predictions, and totals over all sites. Note that these are simple summaries of predictive performance, and make no assumptions about how the data were generated; they are global summary statistics to help evaluate and compare methods. Let $y_{j|r}$ be the true, known value for the jth location and the rth simulation or resampling. Note that if we use $y_{j|r}(\mathbf{s})$ to denote the spatial location $\mathbf{s}$ associated with $y_{j|r}$, then $\{y_{j|r}(\mathbf{s}) \in \mathbb{U}\}$ will vary over $\mathbf{s}$ for each simulation/resampling because observed samples were randomized each time. We indicate this dependence on r as $j|r$. A prediction of $y_{j|r}$, using (1), is denoted $\hat{y}_{j|r}$ for each j and r. For each r, there is a true total T_r, and an estimate, using (2), denoted $\hat{T}_r$ for the rth simulation/resampling. The three performance measures are:

- RMSPE: the root-mean-squared prediction error measures how close the estimates are to the true values.

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

for point-wise predictions, where R = number simulations (1500) of artificial data or resamplings (500) of real forestry data for point prediction, and m = number of point predictions (300 for artificial data and 1500 for real forestry data) per replication r, and

$$\mathrm{RMSPE}_T = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{T}_r - T_r)^2},$$

for total prediction. A smaller value of RMSPE indicates predictors are closer to true values.

- SRB: signed relative bias. Absolute bias is meaningless, so it is expressed as a fraction of the variability. It is well-known that $\mathrm{MSPE} = \mathrm{bias}^2 + \mathrm{variance}$, and we use MSPE as RMSPE squared from above. The sign of bias is also informative, so signed relative bias as a fraction of variability is computed as,

$$\mathrm{SRB}_\kappa = \mathrm{sign}(\tau_\kappa)\sqrt{\frac{\tau_\kappa^2}{\mathrm{MSPE}_\kappa - \tau_\kappa^2}},$$

where

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

for point-wise predictions, and

$$\tau_T = \frac{1}{R}\sum_{r=1}^{R}(\hat{T}_r - T_r),$$

for total predictions, sign(τ_κ) is the sign (positive or negative) of τ_κ, and $\kappa = y$ for a point-wise performance measure or $\kappa = T$ for a total performance measure. A smaller absolute value of SRB has smaller bias, and a negative sign indicates under-prediction and a positive sign indicates over-prediction.

- PIC90:90% prediction interval coverage, measures how well uncertainty is being estimated. For many predicted values, or over many simulations, a prediction interval should cover the true value with the claimed proportion. For point-wise predictions, the empirical prediction interval coverage was computed as,

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

where $\widehat{\text{se}}(\hat{y}_{j|r})$ is the estimated standard error of $\hat{y}_{j|r}$, taken from (4) for k-NN methods, and from the square root of (9) for the SLM (EBLUP), with covariance parameters estimated from (13). PIC90$_y$ should be near 0.90 if prediction intervals are properly estimated. It is also possible to compute PIC95$_y$ by replacing 1.645 with 1.96 in the formula above, and PIC95$_y$ should be near 0.95. For total predictions,

$$\text{PIC90}_T =$$
$$\frac{1}{R}\sum_{r=1}^{R} I\left(\left(\hat{T}_r - 1.645\widehat{\text{se}}(\hat{T}_r)\right) < T_r \ \& \ T_r < \left(\hat{T}_r + 1.645\widehat{\text{se}}(\hat{T}_r)\right)\right),$$

where $\widehat{\text{se}}(\hat{T}_r)$ is the estimated standard error of $\hat{T}_r$, taken from (5) for k-NN methods, and from the square root of (11) for the SLM (EBLUP), with covariance parameters estimated from (13).

Prediction Methods

Seven prediction methods were examined; five different k-NN methods, multiple regression (a special case of a SLM that assumes independence), and a SLM:

- MAH1: k-NN that uses Mahalanobis distance with $k = 1$.
- MAH5: k-NN that uses Mahalanobis distance with $k = 5$.
- MSN1: k-NN that uses most significant neighbor (MSN) with $k = 1$.
- MSN5: k-NN that uses MSN with $k = 5$.
- bstNN: k-NN that uses both Mahalanobis distance and MSN, and tries $k = 1,2,\ldots,30$, and then chooses the distance matrix

and k with the smallest cross-validation RMSPE from the observed data.

- SLM: a spatial linear model using the same covariates as all k-NN methods as main effects only, with an exponential autocovariance model estimated by REML, and using prediction and variance equations as described in the Review of SLM section.
- LM: multiple regression like SLM but assuming all random errors are independent.

Results

The performance measures for the first set of Gaussian simulated data are presented in Table 1. Note that this table is based on 2000 simulations with 300 predictions per simulation, so 2000 total values were estimated and 600,000 points. As expected, the SLM had the lowest RMSPE, for both point and total predictions. Not only was it lowest, it was dramatically lower than any other predictor. The data were simulated with a high amount of autocorrelation, so this demonstrates how much better SLM can be in that case. When compared to MAH5 and MSN1 (the two commonly-used k-NN methods), SLM reduced RMSPE by 52.6 and 64.1% for the point estimates and 43.1 and 66.8% for the total estimates. SLM was also noticeably better than LM (linear model assuming independence), with reduced RMSPE of 34.8 and 31.8% for point and block prediction, respectively. Among the k-NN methods, MSN5 was best for both points and totals, but still not as good as LM. All methods were essentially unbiased for both points and totals. For all point estimates, prediction interval coverage was near 0.90, as they should be. For total estimates, it appears the MAH1 is a bit too high, and perhaps MAH5 a bit too low.

The performance measures for the second set of Poisson simulated data are presented in Table 1. The SLM again had the lowest RMSPE, for both point and total prediction. SLM reduced RMSPE by 31.3 and 14.8% for point estimates when compared to MAH5 and MSN1, and SLM reduced RMSPE by 23.6 and 23.5% for total estimates. All of the methods appear to be unbiased for point prediction, with generally valid confidence interval coverage. There appears to be some bias among the k-NN methods for predicting the total, and some prediction intervals fall below 0.85 for the k-NN methods. Also, the 0.86 prediction interval coverage for the SLM was a bit low, and this simulation was its poorest performance on that measure.

For predicting a binary variable, we replaced the RMSPE with percent correctly classified (PCC) for point prediction. Only the k-NN methods with $k = 1$ truly predicted values that were 0 or 1, so for all other methods predictions were rounded to 0 or 1. The performance measures for the binary simulated data are listed in Table 1. In fact, the k-NN with $k = 1$ performed most poorly, with the SLM again best. SLM increased PCC by 12.5 and 10.3% over MAH5 and MSN1, respectively. Point prediction appears unbiased for all methods. Prediction interval coverage is poor for the methods. A total of binary variables is rarely of interest in forestry applications compared to estimating proportions. For this simulation, we used the block mean, which is the estimated proportion for binary data, instead of a total. For block prediction, SLM decreased the RMSPE by 22.9 and 24.4% over MAH5 and MSN1, respectively. There may be some bias for MAH1 and bstNN. Prediction interval coverage is a little low for MAH5 and bstNN.

The performance measures for resampling real PMAI forestry data are presented in Table 2 in rows marked with PM. For point

Table 1. Performance summaries from 2000 simulated spatial data sets.

	Data	P/T	MAH1	MAH5	MSN1	MSN5	bstNN	LM	SLM
RMSPE	S1	P	9.329	7.451	5.379	4.423	4.456	3.892	2.443
SRB	S1	P	−0.006	−0.009	0	−0.004	−0.004	−0.002	0.001
PIC90	S1	P	0.897	0.9	0.887	0.889	0.88	0.896	0.892
RMSPE	S1	T	262.6	289.8	174.3	153.3	154.5	139.3	87.8
SRB	S1	T	−0.058	−0.067	−0.003	−0.034	−0.034	−0.02	0.009
PIC90	S1	T	0.952	0.87	0.914	0.886	0.874	0.88	0.887
RMSPE	S2	P	6.445	5.17	6.428	5.146	5.129	5.185	4.414
SRB	S2	P	−0.024	−0.03	−0.009	−0.019	−0.030	0.000	0.003
PIC90	S2	P	0.907	0.922	0.899	0.906	0.906	0.932	0.917
RMSPE	S2	T	320	295.9	296.3	262.3	272.2	283.1	226.1
SRB	S2	T	−0.137	−0.188	−0.047	−0.135	−0.182	−0.033	−0.005
PIC90	S2	T	0.912	0.842	0.9	0.867	0.83	0.86	0.858
PCC	S3	P	0.731	0.767	0.749	0.785	0.795	0.799	0.846
SRB	S3	P	0.009	0.013	0.001	0.002	0.010	0.002	0.002
PIC90	S3	P	0.767	0.884	0.764	0.85	0.844	0.905	0.889
RMSPE	S3	T	0.0395	0.0394	0.0387	0.0334	0.0343	0.0329	0.0298
SRB	S3	T	0.072	0.09	0.003	0.019	0.079	0.018	0.014
PIC90	S3	T	0.919	0.841	0.913	0.882	0.84	0.886	0.884

In the Data column, S1 indicates data from the first simulation method, S2 indicates data from the second simulation method (count data), and S3 indicates data from the third simulation method (binary data), as described in Section "Simulation of Artificial Data." Each data set used 100 samples per simulation, indicated by P in the P/T column, and summaries were based on 300 predictions per resampling, which were then averaged over the 2000 simulations. There was one total estimate per simulation, which were summarized over the 2000 simulations, and indicated by T in the P/T column. Different prediction methods form the rest of the columns and are described in Section "Prediction Methods." Performance measures form the rows and are described in Section "Performance Measures;" however, note that percent correctly classified "PCC" replace RMSPE for point predictions of the binary (S3) simulated data.
doi:10.1371/journal.pone.0059129.t001

prediction, SLM reduced RMSPE by 9.0 and 34.4% over MAH5 and MSN1, respectively. Point prediction appears unbiased for all methods. Prediction interval coverage is quite good for all methods. For predicting a total, there appears to be some bias for k-NN methods using Mahalanobis distance, and prediction intervals are too large for MSN1 and too short for MAH5. The SLM reduces the RMSPE for predicting a total by 21.8 and 25.9% over MAH5 and MSN1, respectively.

The performance measures for resampling real DRYBIOT forestry data are presented in Table 2 in rows marked with DB. For point prediction, the bestNN approached SLM for the smallest RMSPE. The MAH5 method also did quite well, but MSN1 was very poor. For predicting a total, SLM again has the lowest RMSPE. There appears to be some bias for the bestNN method. All prediction intervals are within ± 5% of 90%.

Table 2, in rows marked UN, presents the performance measures for resampling real PMAI forestry data with spatially unbalanced sampling, as shown in Figure 2. For point prediction, this creates substantially more bias than the k-NN methods in Table 2. SLM remains relatively unbiased, again with the smallest RMSPE and valid prediction intervals. For predicting a total, there are large biases for k-NN methods and prediction intervals are far from the nominal 90%. The large bias cause the RMSPE for SLM to be much lower than any k-NN methods.

Most methods showed little bias globally, with generally valid prediction intervals. Yet, the SLM, and geostatistics in general, aims to make prediction intervals that vary in space, while the cross-validation approach used for k-NN is constant in space. We re-ran simulation 1 using the iterated variogram (IterVar) variance estimator of [23] in (18), testing its global and point-wise efficacy,

compared to the SLM predictor and interval, and compared to the k-NN predictor under the SLM model (19), which we label kNNGeo. A scatter-plot of a single simulation, with 300 predictions for the unsampled locations, is shown in Figure 3, which plots $|\hat{y}_i - y_i|$ on the x-axis and $\widehat{s.e.}(\hat{y}_i)$ based on (9), (18), and (19) on the y-axis. We computed Kendall's rank correlation between the true absolute error $|\hat{y}_i - y_i|$ and the estimated prediction standard error, $\widehat{s.e.}(\hat{y}_i)$ for each method. These correlations were computed for 1000 simulations, and then all correlations were plotted as violin plots for each method, which is shown in Figure 4.

Figure 4 shows that, indeed, the individual prediction intervals for the SLM are generally related to the actual errors. In contrast, Figure 4 shows that the IterVar method has no relationship between the prediction intervals and the actual absolute errors. Also, [23] claim that the algorithm is expected to rapidly converge. In our implementation, it converged only 57.4% of the time. It diverged before 30 iterations about 2% of the time. Globally, the IterVar 90% prediction interval has 88.8% coverage. The kNNGeo method of Figure 4 showed the strongest correlation between actual absolute errors and prediction intervals, largely due to the fact that it correctly estimated a dominant component of the error, which was the bias-squared. Using the RMSPE of the kNNGeo method for a 90% prediction interval had 94.2% coverage.

Discussion

This article set out to compare k-NN to the SLM for forestry mapping, and for the estimation of totals or averages of forest resources. In the introduction, we laid out arguments that favor

Table 2. Performance summaries for 500 resamplings of forest data.

	Data	P/T	MAH1	MAH5	MSN1	MSN5	bstNN	LM	SLM
RMSPE	PM	P	2.998	2.371	3.243	2.53	2.362	2.399	2.127
SRB	PM	P	0.02	0.038	0.004	−0.001	0.026	0.004	0.003
PIC90	PM	P	0.895	0.902	0.888	0.894	0.898	0.897	0.899
RMSPE	PM	T	219.1	230.7	243.3	200.9	223.2	197	180.4
SRB	PM	T	0.437	0.712	0.064	−0.019	0.446	0.082	0.058
PIC90	PM	T	0.944	0.838	0.948	0.922	0.834	0.904	0.904
RMSPE	DB	P	90.8	71.3	95	73.8	68.4	69.2	67.3
SRB	DB	P	−0.002	0.000	0.000	0.002	−0.018	0.004	0.005
PIC90	DB	P	0.899	0.903	0.892	0.904	0.912	0.919	0.914
RMSPE	DB	T	6795	6369	7683	6393	6498	6193	6091
SRB	DB	T	−0.027	−0.001	0.027	0.036	−0.302	0.052	0.066
PIC90	DB	T	0.942	0.878	0.914	0.878	0.848	0.876	0.866
RMSPE	UN	P	2.983	2.497	3.115	2.495	2.389	2.436	2.146
SRB	UN	P	0.135	0.227	0.08	0.104	0.139	0.159	0.028
PIC90	UN	P	0.912	0.907	0.905	0.903	0.9	0.903	0.918
RMSPE	UN	T	637.9	853.6	457.1	442.2	576.1	608.1	269
SRB	UN	T	2.635	4.055	1.418	1.77	1.651	2.86	0.369
PIC90	UN	T	0.248	0.01	0.626	0.438	0.308	0.128	0.92

In the Data column, PM indicates the PMAI data set, DB indicates the DRYBIOT data set, and UN indicates the PMAI data set with unbalanced sampling, as described in Section "Forest Data." Each data set used 386 samples per resampling, and for point predictions, indicated by P in the P/T column, summaries were based on 1500 predictions per resampling, which were then averaged over the 500 resamples. There was one total estimate per resample, which were summarized over the 500 resamples, and indicated by T in the P/T column. Different prediction methods form the rest of the columns and are described in Section "Prediction Methods." Performance measures form the rows and are described in Section "Performance Measures."
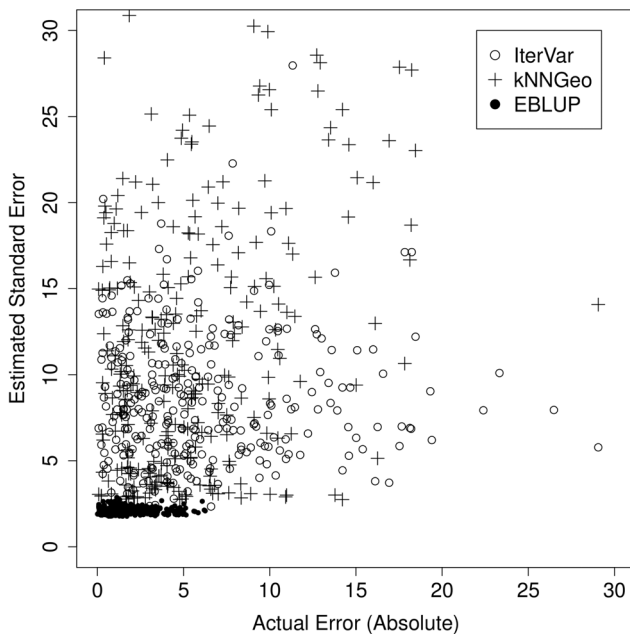doi:10.1371/journal.pone.0059129.t002

Figure 3. Scatter plots of absolute errors and the estimated standard errors for a single simulated data set. IterVar is the iterated variogram method of McRoberts et al. (2007), kNNGeo is the covariance matrix as estimated with all main effects in a spatial linear model and REML, but using the k-NN weights, and EBLUP are the estimated standard errors from the SLM.
doi:10.1371/journal.pone.0059129.g003

using k-NN, and arguments that favor using a SLM, along with disadvantages for both. Our simulations of artificial data and resamplings of real data are not exhaustive; however, for the criteria that we chose (RMSPE, signed relative bias, and prediction interval coverage), the results presented in the previous section clearly favor SLM in general. To summarize, we simulated data under conditions that should severely test the SLM method. Because k-NN is primarily used in forestry, we included various k-NN methods in the simulations. In all cases, even with mis-specified covariance models, mis-specified linear models (including nonsignificant covariates and ignoring significant ones), zero-inflated count data, binary data, and skewed real forestry data, the SLM performed better than k-NN, and generally provided valid inference with little bias, and prediction intervals that contained the true values the correct proportion of time. From a single simulation, it also appears that the SLM is more robust to unbalanced spatial sampling. These results generally verify the claim in the introduction that EBLUP used to estimate the SLM is fairly robust in a variety of ways. The SLM has an additional benefit from its model-based assumptions; it allows point-wise inference, with globally valid prediction intervals that vary at each point.

Our results can be compared to previous literature cited in the Introduction, such as [46], where our SLM is mathematically equivalent to their universal kriging (UK); however, parameter estimation likely differed in the studies ([46] do not specify if they used the REML option when they fit variograms using the GSTAT package [56]). In [46], the SLM performed well compared to another k-NN method called gradient nearest neighbor (GNN), but not as consistently better as our results. Our results can also be compared to [47], where our SLM is
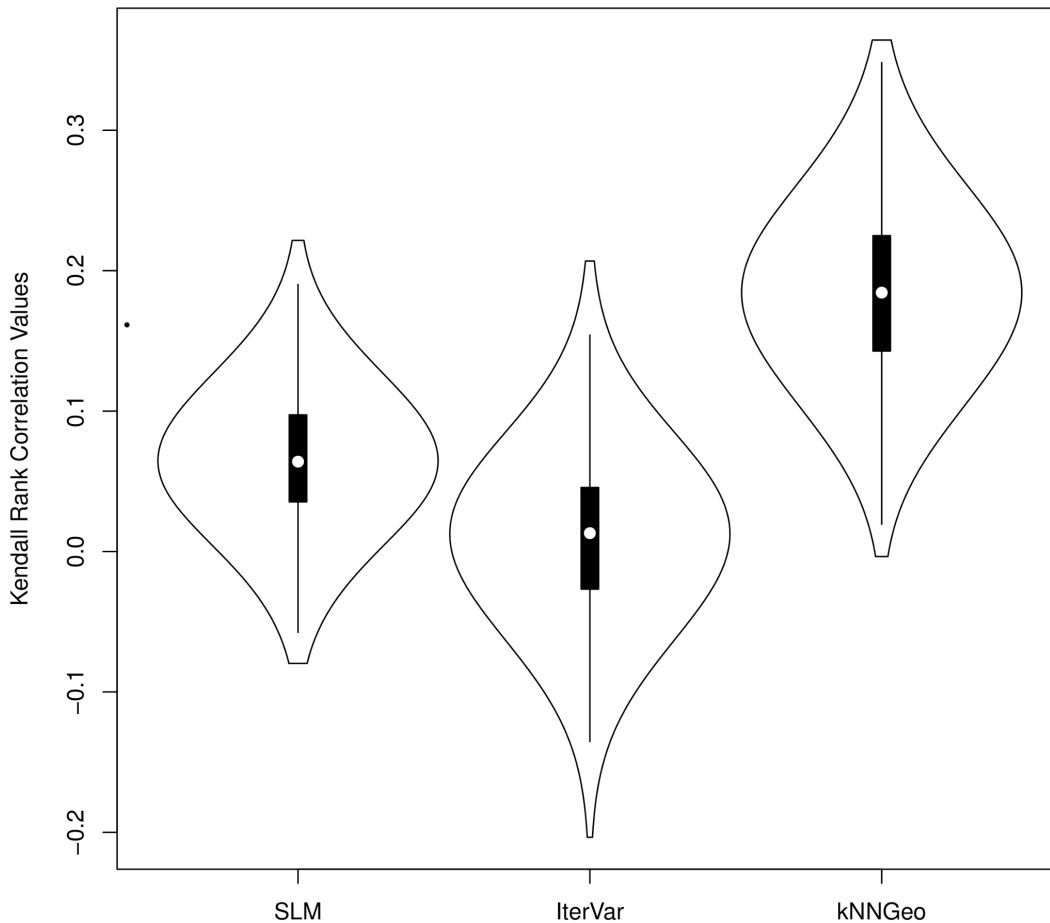
Figure 4. Violin plots of Kendall's rank correlation coefficients between absolute error and estimated standard errors over 2000 simulations.

mathematically equivalent to their kriging-with-external-drift (KED). An interesting hybrid method that uses MSN with kriging on the residuals is compared to the SLM based on RMSPE and bias [47]. They find that the MSN-kriging hybrid is slightly better than the SLM, with both better than MSN alone. However, we note that [47] do not give a standard error estimator of point-wise predictions for the MSN-kriging hybrid. Also [47] estimate the SLM by first fitting a linear model assuming independence, and then computing and fitting a semivariogram on residuals. The use of REML for the SLM, as we have described it, estimates the fixed effects assuming correlated residuals, and is expected to be more efficient.

Note that it may seem surprising that k-NN was mostly unbiased for these simulations. Clarification is required because the Introduction claimed that k-NN is biased [16–19]. These authors equate bias with the fact that k-NN underestimates large values and overestimates small values; in geostatistics, this characteristic is called smoothing [51](pg. 158). Smoothing is a desirable characteristic under squared-error loss, which the SLM minimizes, so it is also a property of the SLM [57](pg. 189). Because the SLM is BLUP, it is unbiased for point-wise predictions; however the predictions are not unbiased for nonlinear functionals of the spatial population, such as quantiles. For example, following [58], but for finite populations, $F(y; \{Y_i\}) = (n+m)^{-1} \sum_i^{n+m} I(Y_i < y)$ is a spatial cumulative distribution function (SCDF). Then its inverse, $F^{-1}(\alpha; \{Y_i\}) \equiv \arg\min[y \in \mathbb{R}; F(y; \{Y_i\}) > \alpha]$ defines the α quan-

tile, which is nonlinear in $\{Y_i\}$. Predictors that can handle such nonlinearity have been proposed [59], and by matching variances in the predictions to those in the data, predictions will no longer underestimate high values or overestimate low values. However, it should be noted that these predictors will sacrifice the pointwise MSPE as optimized for BLUP; for an example where the prediction-variance-constrained MSPE is twice that of the "smooth" SLM predictor, see [59]. This illustrates that, in general, no set of predictors will be optimal for all purposes.

More generally, it is possible, though computationally expensive, to obtain multiple sets of predictions, where the predicted data are simulated from conditional distribution properties of the population. The multiple prediction sets can be averaged to obtain predictions that satisfy BLUP, or quantiles can be computed across the sets. Multiple sets of predictions also allow the propogation of uncertainty if prediction sets are used as inputs to other models. In fact, k-NN is closely related to multiple imputation methods [60–63], which sample from existing data to impute for missing data; in that sense they are like using $k=1$ multiple times to give multiple possible "realizations." Again, there are equivalent ideas in geostatistics, generally termed "conditional simulation," e.g., [64] and [51](pgs. 452–453). We do not pursue a comparison here but suggest it for further research. Given the above discussion, we emphasize that our goal was point-wise unbiased prediction while minimizing the MSPE, which is what the SLM

achieves in a single map, and compare that to k-NN on the same basis.

A model-based variance for k-NN predictors remains problematic. Cross-validation works from a global standpoint. Attempts at making variance local, such as the iterated variogram approach [23], did not work well for one simulated data set as shown by Figure 4. There was no correlation between estimated prediction standard errors and actual absolute errors, so cross-validation was just as good, and the iterated variogram approach had convergence problems and does not work with $k = 1$. More testing of this method, and possible improvements are warranted. The kNNGeo method was correlated to actual absolute errors. However global prediction intervals for kNNGeo were too conservative with 94.2% coverage for 90% prediction intervals because the bias component is not stochastic, but is treated as such if included in prediction intervals. Thus, the SLM is the only viable choice that we examined for making valid uncertainty maps along with predictions.

Finally, we stress that the SLM was presented here as a "black box" method. As we used it, there were no decisions involved; after choosing a covariance model like the exponential, use all covariates that are available, estimate the covariance parameters with REML, and plug the resulting covariance matrix into the prediction equations. This allowed us to make predictions for thousands of simulations. Such a "black box" method is certainly possible when many predictions are needed by personnel with little statistical training. However, when data have been collected at great expense, a careful analysis is better. In that case, exploratory data analysis, understanding of covariate relationships, finding and explaining outliers, model selection and diagnostics, and finally prediction, all can enhance prediction and understanding for both the SLM and k-NN, and we recommend that over a black box approach. For example, a Bayesian approach can account for the fact that covariance parameters are estimated and should correct for the plug-in aspect of the EBLUP, e.g. [43,65], with available software [66]. Also, when remotely sensed data are involved, data sets can be massive in size. In that case, other methods can be used, e.g. [67,68]. Moreover, there is no single correct analysis for forestry data sets; they can be modeled in various ways to achieve different desired goals.

Author Contributions

Conceived and designed the experiments: JVH HT. Performed the experiments: JVH. Analyzed the data: JVH. Wrote the paper: JVH HT.

References

1. Moeur M, Stage AR (1995) Most similar neighbor: An improved sampling inference procedure for natural resource planning (STMA V37 0534). Forest Science 41: 337–359.
2. Haara AM, Maltamo M, Tokola T (1997) The k-nearest neighbor methods for estimating basal area diameter distribution. Scandinavian Journal of Forest Research 12: 200–208.
3. LeMay V, Temesgen H (2005) Comparison of nearest neighbor methods for estimating basal area and stems per hectare using aerial auxiliary variables. Forest Science 51: 109–199.
4. Holmström H, Fransson J (2003) Combining remotely sensed optical and radar data in knnestimation of forest variables. Forest Science 16: 409–418.
5. McRoberts R (2008) Using satellite imagery and the k-nearest neighbors technique as a bridge between strategic and management forest inventories. Remote Sensing of Environment 112: 2212–2221.
6. Tuominen S, Fish S, Poso S (2003) Combining remote sensing, data from earlier inventories, and geostatistical interpolation in multisource forest inventory. Canadian Journal of Forest Research 33: 624–634.
7. Tomppo E, Gagliano C, De Natale F, Katila M, McRoberts R (2009) Predicting categorical forest variables using an improved k-nearest neighbor and landsat imagery. Remote Sensing of Environment 113: 500–517.
8. Temesgen H, LeMay VM, Froese KL, Marshall PL (2003) Imputing tree-lists from aerial attributes for complex stands of south-eastern british columbia. Forest Ecology and Management 177: 277–285.
9. Katila M, Tomppo E (2001) Selecting estimation parameters for the finnish multisource national forest inventory. Remote Sensing of Environment 76: 16–32.
10. Fehrmann L, Lehtonen A, Kleinn C, Tomppo E (2008) Comparison of linear and mixed-effect regression models and a k-nearest neighbour approach for estimation of single-tree biomass. Canadian Journal Forest Research 38: 1–9.
11. Sironen S, Kangas A, Maltamo M, Kalliovirta J (2003) Estimating individual tree growth with nonparametric methods. Canadian Journal of Forest Research 33: 444–449.
12. Maltamo M, Malinen J, Kangas A, Härkönen S, Pasanen AM (2003) Most similar neighbour-based stand variable estimation for use in inventory by compartments in finland. Forestry 76: 449–464.
13. Malinen J (2003) Locally adaptable non-parametric methods for estimating stand characteristics for wood procurement planning. Silva Fenn 37: 109–118.
14. Sironen S, Kangas A, Maltamo M, Kangas J (2008) Localization of growth estimates using nonparametric imputation methods. Forest Ecology and Management 256: 674–684.
15. Crookston NL, Finley AO (2008) yaimpute: An R Package for kNN Imputation. Journal of Statistical Software 23: 1–16.
16. McRoberts R, Nelson M, Wendt D (2002) Stratified estimation of forest area using satellite imagery, inventory data, and the k-nearest neighbors technique. Remote Sensing of Environment 82: 457–468.
17. Baffetta F, Fattorini L, Franceschi S, Corona P (2009) Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. Remote Sensing of Environment 113: 463–475.
18. Packalén P, Maltamo M (2007) The k-msn method for the prediction of species-specific stand attributes using airborne laser scanning and aerial photographs. Remote Sensing of the Environment 109: 328–341.
19. Baffetta F, Corona P, Fattorini L (2012) A matching procedure to improve k-nn estimation of forest attribute maps. Forest Ecology and Management 272: 35–50.
20. Stage A, Crookston N (2007) Partitioning error components for accuracy-assessment of nearneighbor methods of imputation. Forest Science 53: 62–72.
21. McRoberts R (2009) Diagnostic tools for nearest neighbors techniques when used with satellite imagery. Remote Sensing of Environment 113: 489–499.
22. Kim HJ, Tomppo E (2006) Model-based prediction error uncertainty estimation for k-nn method. Remote Sensing of Environment 104: 257–263.
23. McRoberts R, Tomppo EO, Finley AO, Heikkinen J (2007) Estimating areal means and variances of forest attributes using the k-nearest neighbors technique and satellite imagery. Remote Sensing of Environment 111: 466–480.
24. Magnussen S, McRoberts RE, Tomppo EO (2009) Model-based mean square error estimators for k-nearest neighbour predictions and applications using remotely sensed data for forest inventories. Remote Sensing of Environment 113: 476–488.
25. Cressie N (1990) The origins of kriging. Mathematical Geology 22: 239–252.
26. Cressie NAC (1993) Statistics for Spatial Data. New York: John Wiley and Sons. 900 p.
27. Ver Hoef JM (2000) Predicting finite populations from spatially correlated data. In: ASA Proceedings of the Section on Statistics and the Environment. American Statistical Association, 93–98.
28. Ver Hoef JM (2002) Sampling and geostatistics for spatial data. Ecoscience 9: 152–161.
29. Ver Hoef JM (2008) Spatial methods for plot-based sampling of wildlife populations. Environmental and Ecological Statistics 15: 3–13.
30. Stein ML (1988) Asymptotically efficient prediction of a random field with a misspecified covariance function. The Annals of Statistics 16: 55–63.
31. Putter H, Young GA (2001) On the effect of covariance function estimation on the accuracy of kriging predictors. Bernoulli 7: 421–438.
32. Cressie N (1985) Fitting models by weighted least squares. Journal of the International Association for Mathematical Geology 17: 563–586.
33. Patterson HD, Thompson R (1971) Recovery of inter-block information when block sizes are unequal. Biometrika 58: 545–554.
34. Harville DA (1977) Maximum likelihood approaches to variance component estimation and to related problems (C/R: P338–340). Journal of the American Statistical Association 72: 320–338.
35. Heyde CC (1994) A quasi-likelihood approach to the REML estimating equations. Statistics & Probability Letters 21: 381–384.
36. Cressie N, Lahiri SN (1996) Asymptotics for REML estimation of spatial covariance parameters. Journal of Statistical Planning and Inference 50: 327–341.
37. Bolthausen E (1982) On the central limit theorem for stationary mixing random fields. The Annals of Probability 10: 1047–1050.
38. Cressie N, Zimmerman DL (1992) On the stability of the geostatistical method. Mathematical Geology 24: 45–59.

39. Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics (Disc: P326–350). Journal of the Royal Statistical Society, Series C: Applied Statistics 47: 299–326.
40. Zimmerman DL, Cressie N (1992) Mean squared prediction error in the spatial linear model with estimated covariance parameters. Annals of the Institute of Statistical Mathematics 44: 27–43.
41. Ver Hoef J, Cressie N (1993) Multivariable spatial prediction. Mathematical Geology 25: 219–240.
42. Heisel T, Ersbll AK, Andreasen C (1999) Weed mapping with co-kriging using soil properties. Precision Agriculture 1: 39–52.
43. Finley AO, Banerjee S, Ek AR, McRoberts R (2008) Bayesian multivariate process modeling for prediction of forest attributes. Journal of Agricultural, Biological, and Environmental Statistics 13: 60–83.
44. Finley AO, Banerjee S, McRoberts RE (2009) Hierarchical spatial models for predicting tree species assemblages across large domains. The Annals of Applied Statistics 3: 1052–1079.
45. Finley AO, McRoberts RE (2008) Efficient k-nearest neighbor searches for multi-source forest attribute mapping. Remote Sensing of Environment 112: 2203–2211.
46. Pierce KB Jr, Ohmann JL, Wimberly MC, Gregory MJ, Fried JS (2009) Mapping wildland fuels and forest structure for land management: a comparison of nearest neighbor imputation and other methods. Canadian Journal of Forestry Research 39: 1901–1916.
47. Räty M, Kangas A (2012) Comparison of k-msn and kriging in local prediction. Forest Ecology and Management 263: 47–56.
48. Schabenberger O, Gotway CA (2005) Statistical Methods for Spatial Data Analysis. Boca Raton, FL: Chapman Hall/CRC. 512 p.
49. R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3–900051–07–0. Available: http://www.R-project.org. Accessed 2013 Feb 25.
50. Thompson SK (1992) Sampling. New York: John Wiley and Sons. 343 p.
51. Chiles JP, Delfiner P (1999) Geostatistics: Modeling Spatial Uncertainty. New York: John Wileyand Sons. 695 p.
52. Nelder JA, Mead R (1965) A simplex method for function minimization. Computer Journal 7: 308–313.
53. Roesch F, Reams G (1999) Analytical alternatives for an annual inventory system. Journal of Forestry 97: 44–48.
54. Czaplewski R (1999) Forest survey sampling designs: a history. Journal of Forestry 97: 4–10.
55. Wang T, Hamann A, Spittlehouse D, Aitken S (2006) Development of scale-free climate data for western canada for use in resource management. International Journal of Climatology 26: 383–397.
56. Pebesma EJ (2004) Multivariable geostatistics in s: the gstat package. Computers & Geosciences 30: 683–691.
57. Johnston K, Ver Hoef J, Krivoruchko K, Lucas N (2001) Using ArcGIS geostatistical analyst, volume 300. Redlands, CA: ESRI Press. 300 p.
58. Lahiri SN, Kaiser MS, Cressie N, Hsu NJ (1999) Prediction of spatial cumulative distribution functions using subsampling (C/R: P97–110). Journal of the American Statistical Association 94: 86–97.
59. Aldworth J, Cressie N (2003) Prediction of nonlinear spatial functionals. Journal of Statistical Planning and Inference 112: 3–41.
60. Schafer JL, Olsen MK (1998) Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research 33: 545–571.
61. Schafer JL (1999) Multiple imputation: A primer. Statistical Methods in Medical Research 8: 3–15.
62. Rubin DB (1996) Multiple imputation after 18+ years. Journal of the American Statistical Association 91: 473–489.
63. Rubin DB (1996) Reply to comments on \Multiple imputation after 18+ years. Journal of the American Statistical Association 91: 515–517.
64. Ravenscroft PJ (1994) Conditional simulation for mining: Practical implementation in an industrial environment (Disc: P106–109). In: Armstrong M, Dowd PA, editors. Geostatistical Simulations. New York: Kluwer Academic Publishers Group. 79–87.
65. Handcock MS, Stein ML (1993) A Bayesian analysis of kriging. Technometrics 35: 403–410.
66. Finley AO, Banerjee S, Carlin BP (2007) spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. Journal of Statistical Software 19: 1–24.
67. Cressie N, Johannesson G (2008) Fixed rank kriging for very large spatial data sets. Journal of the Royal Statistical Society, Series B 70: 209–226.
68. Eidsvik J, Finley AO, Banerjee S, Rue H (2012) Approximate bayesian inference for large spatial datasets using predictive process models. Computational Statistics and Data Analysis 56: 1362–1380.