REGRESSION KRIGING HAS MARGINAL IMPACT ON LANDSAT-BASED PREDICTIONS OF FOREST BIOMASS ACROSS NEW YORK STATE

Lucas Johnson
Ph.D. Student
Graduate Program in Environmental Science,
State University of New York College of Environmental Science and Forestry,
ljohns11@esf.edu

Abstract

There is a growing need for spatially explicit, high-resolution, biomass predictions as federal, state, and international agencies look to forests to offset economy-wide emissions. Optical satellite imagery, like Landsat, offers historically consistent data with wide spatial coverage, and has been relied upon as the primary information source in many biomass mapping and modeling efforts. However, Landsat based models are not without significant errors. In this report, I attempted to leverage the spatial relationships between Landsat-based model residuals at federally maintained Forest Inventory and Analysis (FIA) plot locations to improve model predictions in New York State. I used variogram models and ordinary kriging to produce predictions of residuals which were then added back to initial model predictions. These improved predictions were assessed against FIA plots and compared to original model predictions at a set of holdout plots. Additionally, I demonstrated how cross-validation paired with a grid search can be an effective method to produce variogram parameters. Neither manually fit variograms, nor cross-validation fit variograms were able to meaningfully improve model predictions at FIA plot locations.

Introduction

Forest mapping and monitoring is becoming increasingly important as federal, state, and global agencies look towards natural solutions to mitigate a warming climate and the myriad resulting challenges. Field sampling programs, like the United States Department of Agriculture's Forest Inventory and Analysis program (FIA) (Gray et al. 2012), provide unbiased estimates of forest structure over large areas, but lack the fine spatial resolution to understand and manage forests at relevant scales. Thus, high-resolution forest mapping is needed to inform decision-makers where forest resources should be managed or preserved. The New York State (NYS) legislature in particular has mandated that the state reach net-zero emissions across the entire economy by the year 2050. To this end NYS is actively researching the potential carbon benefits that their forests, which dominate 60% of the statewide landscape, can offer.

Since the United States Geological Survey (USGS) opened the Landsat data archive in 2008, there has been an explosion of terrestrial monitoring approaches built with these freely available global datasets (Wulder et al. 2012; Banskota et al. 2014). Landsat offers the longest history of publicly available remote sensing data, from 1972 to present day, and has repeat observations for the same location roughly twice a month. These data are available everywhere, and Landsat missions are well supported meaning that any methods developed now can likely applied for future monitoring. Given the availability, moderate resolution (30m), temporal density, and global spatial coverage of Landsat data, many researchers have explored their utility for mapping and monitoring

forest conditions over large scales and long time periods. Landsat forestry applications include forest cover or species mapping, disturbance mapping, and biophysical (e.g. basal area, biomass, and canopy height) mapping (Banskota et al. 2014).

While the benefits of Landsat imagery in terms of coverage and availability are well documented, it is also well known that they cannot predict forest structure as accurately as active remote sensing data like LiDAR (Huang et al. 2019; Hurtt et al. 2019; Chen and McRoberts 2016). Landsat-based models inherently rely on measures of spectral reflectance or 'greenness' to predict forest structure. This is challenging in that vegetation near the ground floor can and does appear quite green, and at a certain point more mature forests reach a greenness saturation point after which biomass, basal area, or height growth may still proceed without changes in spectral reflectance. This inherent limitation in Landsat's ability to model forest structure leads to less accurate spatial predictions of biomass and carbon and this can have large impacts on greenhouse gas budgets that rely these mapped predictions. Landsat-based model predictions have been shown to be more accurate when aggregated to larger scales (Riemann et al. 2010), however accurate predictions at smaller scales will open the door to more targeted local management (e.g. private land parcel, individual forest stands) for improved forest carbon outcomes.

Regression kriging, or kriging of model residuals, is an approach that has been previously documented to improve spatial predictions of forest structure which rely on both optical and radar satellite imagery (A. Hudak et al. 2002; Meng, Cieszewski, and Madden 2009; Tsui et al. 2013). In this report I assessed the benefit of regression kriging to improve Landsat-based model predictions of forest aboveground biomass (AGB) across New York State (NYS). I examined first order effects effects of model residuals, fit variograms to the residuals, and spatially interpolated model residuals using ordinary kriging. I created enhanced predictions at a set of holdout FIA plots by adding the kriged residual prediction to the original model prediction. More accurate spatial predictions of AGB can help NYS target areas for improved forest management with the goal of improved carbon sequestration and avoided carbon emissions to help achieve net-zero emissions in NYS by 2050.

Literature Review

Several previous studies have explored the efficacy of spatial interpolation for mapping components of forest structure like mean biomass per unit area, canopy height, and basal area. Two of such studies compared aspatial regression and machine learning (ML) methods to spatial interpolation methods in the form of ordinary kriging, co-kriging, and regression kriging (A. Hudak et al. 2002; Freeman and Moisen 2006). A. Hudak et al. (2002) showed that aspatial methods, models based on Landsat-derived predictors, better maintained the pattern of vegetation across the study area while the spatial methods were less biased when producing maps of forest canopy height in a 200 km^2 research forest in western Oregon. A. Hudak et al. (2002) categorized regression kriging as an 'integrated' approach in that the primary models relying on Landsat information are aspatial, and then the spatial relationships among the model residuals are used to improve the predictions. The integrated approach was better than the strictly spatial or aspatial approaches in that predictions were both unbiased and also matched the spatial pattern of vegetation. Freeman and Moisen (2006) produced similar comparisons between spatial and aspatial methods, but over a much larger study area covering the entire Rocky Mountain region (RMR). This region was broken up into 18 zones based on ecological similarity (Homer and Gallant 2001) to reduce the potential effects of

an unknown trend across the RMR. Despite these efforts, they found that their aspatial model built with MODIS-derived predictors, was much better than the aspatial methods, and regression kriging did little to improve the aspatial model predictions. They attribute these outcomes to the fact that the aspatial model already incorporated much of the spatial pattern in the landscape through the MODIS and environmental predictor layers built into the model.

Additionally, two other studies compared various kriging approaches to each other. Tsui et al. (2013) used synthetic aperture radar (SAR) as auxiliary information to estimate AGB across a 25 km^2 area in Vancouver Island, Canada, while Meng, Cieszewski, and Madden (2009) used Landsat imagery as auxiliary information to estimate pine basal area across 20 counties in Georgia, United States. Tsui et al. (2013) showed that regression kriging was more accurate in terms of root mean squared error (RMSE) and mean absolute error (MAE) than all other kriging approaches (ordinary kriging, co-kriging, regression co-kriging). Meng, Cieszewski, and Madden (2009) showed that regression kriging was superior to ordinary kriging and co-kriging in terms of R^2.

Three out of the four studies here showed that regression kriging produced the best results among spatial and aspatial approaches, however each of these three studies operated on much smaller scales (25 km^2, 250 km^2, ~35,000 km^2) than the ~141,000 km^2 in NYS. The Freeman and Moisen (2006) study, the one study listed here where regression kriging was not the best approach, operated on a much larger area than NYS, but broke the region up into smaller units. Additionally, this study noted that the scale of variation in the mountainous RMR was smaller than the distance between reference FIA plots leading to a high nugget effect for the spatial approaches.

Study Area

The study area for this report was all of NYS (Figure 1), spanning 125,411,232 km^2, and 62 counties. The topography across the state varies from 0m above sea level to roughly 1,650m above sea level in the Adirondack region in the northern portion of the state. Roughly 60% of the state is forested, and this forested land is dominated by the maple, beech, birch forest type (53%). The majority of forest land in NYS is privately owned (76%) (DEC, n.d.).

Datasets

Estimates of AGB for all trees measuring $\geq$ 12.7 cm (5 in) diameter at breast height were produced as part of the USDA FIA program (Gray et al. 2012), with true plot centroid locations obtained under agreement with the USDA. Estimates were recorded in pounds, then summed at each plot and area-normalized to units of megagrams per hectare (Mg ha^{-1}). The plots are sampled on a hexagonal grid, with random offsets within each hexagon, such that one plot is sampled roughly every 2,400 ha (~6000 acres). However, only one fifth of the plots are sampled in each year to reduce sampling costs (Bechtold and Patterson 2005). For the purposes of this study, plots inventoried in 2019, the most recent inventory that is publicly available, were used to limit processing time, and to produce temporally coherent results. This field data was partitioned into a roughly 70% training dataset (426 plots), while the remaining 30% of the plots were set aside as a testing dataset (168 plots). The min, max, and average distances between plots are recorded in Table 1.

I used a modeling approach to produce initial predictions and residuals for this study that closely followed the approach developed by A. T. Hudak et al. (2020). I developed a set of 20 predictors derived from Landsat analysis ready data (Dwyer et al. 2018), Landtrendr derived disturbance
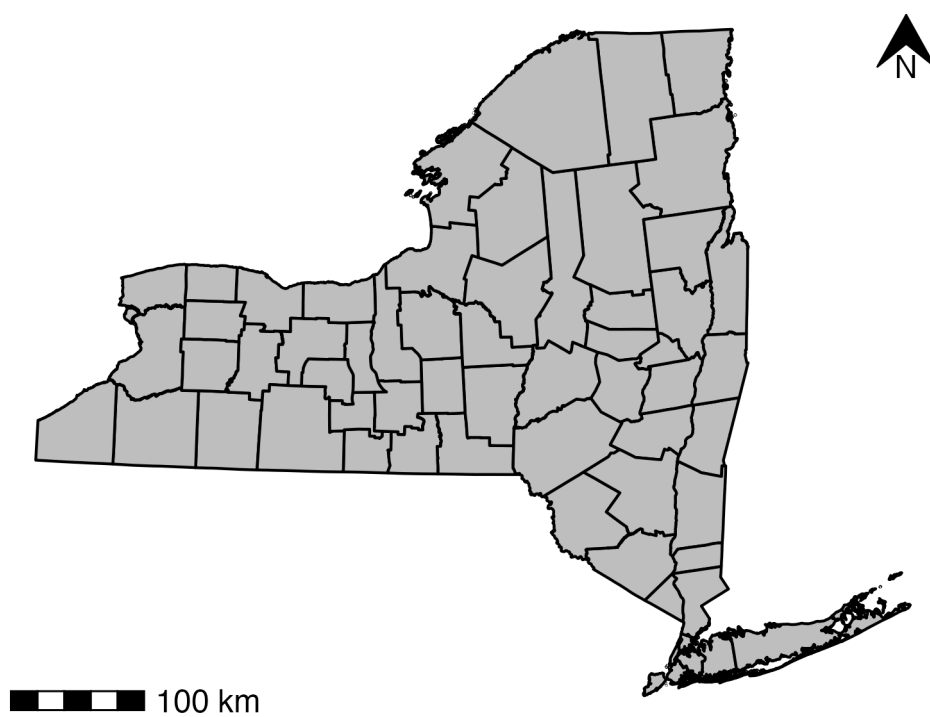
Figure 1: New York State county shoreline map.

Table 1: Plot distance summary in kilometers; Min and Max in km; Intensity in plots per thousand km^2

Partition	Min Distance	Max Distance	Intensity
Train	2.08	631.45	3.01
Test	2.46	608.43	1.19

and temporal segmentation information (Kennedy et al. 2018), a global forest canopy height layer (Simard et al. 2011), topographic data (Mahoney 2021), climate data (Daly et al. 2008), and land cover classifications (Brown et al. 2020; Zhu and Woodcock 2014). I fit three ML models to a randomly selected 70% of the training dataset. Specifically I fit a random forest model (ranger, Wright and Ziegler (2017)), a stochastic gradient boosting machines (lightgbm, Ke et al. (2021)), and a support vector machines (kernlab, Karatzoglou et al. (2004)). With these 3 component models, I developed a "stacked" linear model ensemble model (hereafter LINMOD) in effort to reduce the generalization error of the component models (Wolpert 1992). I developed LINMOD by regressing the component model predictions for the 30% of the training data not used to train the component models against the observed values.

I then used LINMOD to make predictions for 30m pixels across the entire state. To produce plot-level residuals, I summarized pixel predictions at the training plot locations as well as the testing plot locations by taking the weighted average of pixels that intersected each plot. I computed residuals by subtracting reference AGB values from the original model prediction such that positive residuals represented overpredictions and negative residuals represented underpredictions.

Methods and Results

First, I examined the first order effects of the data to identify any global trends or directional patterns in the data. Before attempting to visualize the data spatially, I produced a histogram of the training data residuals (Figure 2). The histogram of the residuals indicated that the data was roughly normally distributed and centered on 0, which is perhaps unsurprising given that the the attributes are model residuals which ideally have a mean of 0. The data is slightly skewed left, with a few large negative residuals near 200 Mg ha^{-1}. For the sake of mapping first order effects, I interpreted these residuals as outliers, and the color scale range was restricted to $[-max(residual), max(residual)]$ such that these large negative outliers were included, but the color scale was not shifted drastically due to their inclusion.

To visualize the first order effects spatially, I plotted a continuous point kernel function which estimates local means where attributes were weighted by their distance from the estimation location (kernel center). The continuous point kernel estimates were computed as follows:

$$\hat{\mu_\tau}(s) = \sum_{i=1}^{n} w_i(s) y_i \tag{1}$$

Where $\hat{\mu_\tau}(s)$ is the average estimate for location s, n is the number of points within a distance τ of location s, w_i is the weight for point i and y_i is the attribute value for point i. w_i is computed as follows:
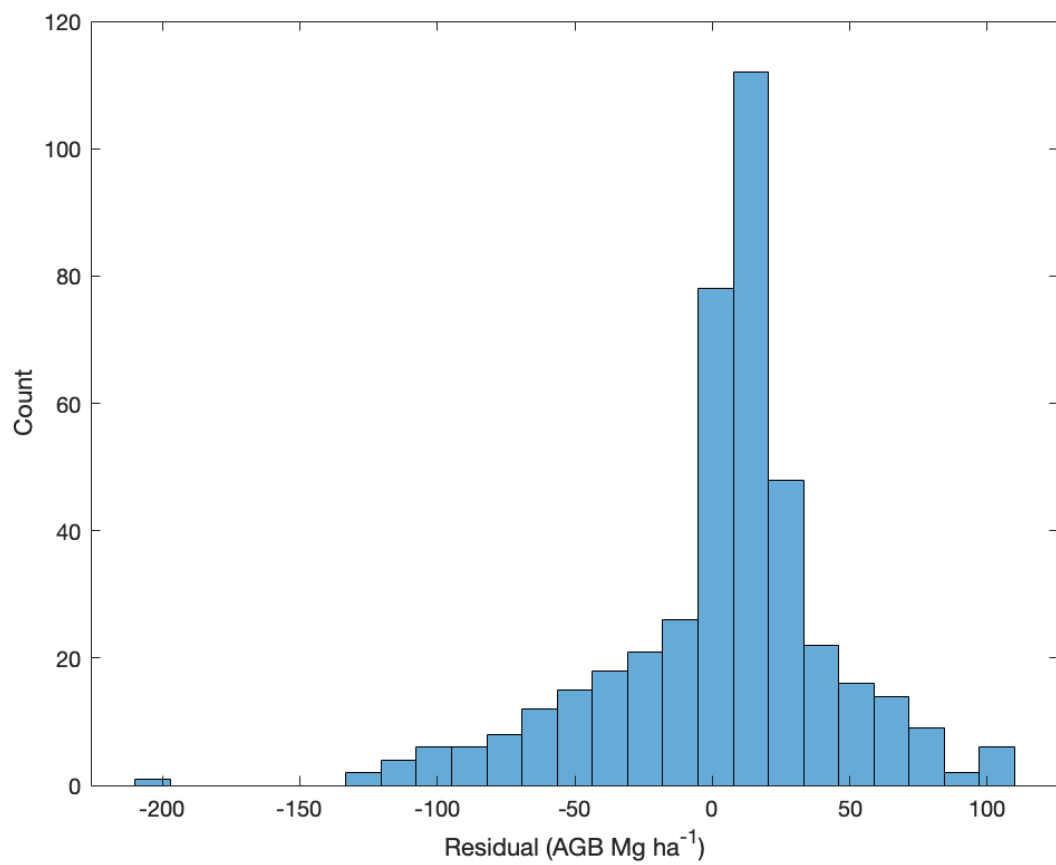
Figure 2: Histogram for initial model residuals at training data locations. 25 bins used.

$$w_i(s) = \frac{k(\frac{(s-s_i)}{\tau})}{\sum_{j=1}^{n} k(\frac{(s-s_j)}{\tau})} \tag{2}$$

Where k is computed as follows:

$$k = \frac{3}{\pi\tau^2}(1 - \frac{h^2}{\tau^2})^2 \tag{3}$$

Where h is the distance between two locations. I chose a bandwidth of 250km and uniformly spaced kernels 5km apart to produce a smooth surface with acceptable resolution. For the color scale I chose equal intervals with $1 + 3.3log(n)$ classes, centered on 0. The range was trimmed as mentioned above.

The plotted continuous point kernel surface (Figure 3) does not show any identifiable trends or directional patterns. Rather, the spatial arrangement of residuals seems random, with underpredictions and overpredictions not following any strong spatial pattern. This can be interpreted in two different ways. First, we might conclude that the environmental and spatial variables used as predictors in the model were able to capture any large scale trends in biomass across the state. Alternatively, if we assume that biomass is "randomly" distributed across the state, and since these residuals are partly a function of the measured biomass on the ground, it would follow that the residuals themselves are randomly distributed. We do know that there are large concentrations of forest in the Adirondack and Catskill regions, however most of the state is dominated by patchy forest cover.

Next, I explored variograms to identify the second order effects of the data with the goal of leveraging the spatial dependence of the residual values for spatial interpolation. Specifically, variograms help identify the way the attribute deviations covary across the area of interest. The results of the continuous point kernel analysis led to the use of a single omnidirectional variogram for the entire state. The variogram estimates for a given lag were computed as follows:

$$\hat{\gamma}(h) = \frac{1}{2n(h)} \sum_{h=s_i-s_j} (y_i - y_j)^2 \tag{4}$$

Where h represents a lag distance between two locations s_i and s_j, n is the number of points that have a distance smaller than lag h, and y_i, y_j are the attribute values at locations i and j respectively.

First, I plotted the variogram estimates using the training data and 700 equally spaced lag distances (Figure 4). Additionally, I plotted the variogram cloud (Figure 5) where instead of the average pairwise relationship at each lag, all individual pairwise relationship were plotted at each lag. I limited the variogram and variogram cloud plots to lags $\leq$ 50 km which omitted a good portion of the statewide data that contains distances up to roughly ~600 km (Table 1). We don't expect forest conditions at one end of the state to have any influence on forest conditions at the other end of the state. Moreover, the forest conditions in NYS are quite patchy which is reflective of the majority privately owned forest lands in the state (*Datasets* section), with small private ownerships creating a parcelized mosaic of management patterns and conditions. This patchiness restricts biomass similarities to smaller scales This is not uncommon in the northeastern United States, but is much different than the West where large swaths of forest are maintained on public lands. The 50 km is likely still generous outside of the Adirondack or Catskill region in the state, but allows us to focus
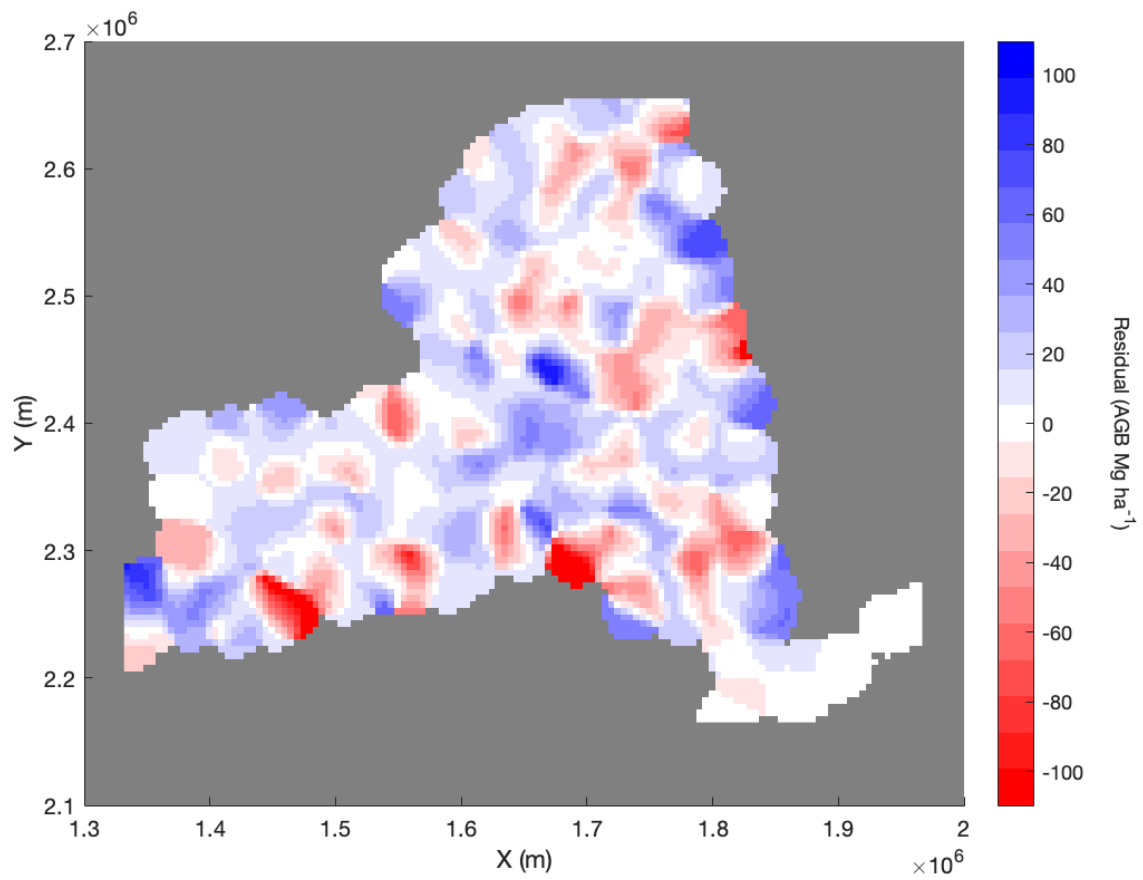
Figure 3: Continuous point kernel estimates for training data; Bandwidth = 25 km; Kernels uniformly spaced 5 km apart.

on the covariance at smaller lag distances. The variogram cloud helps make sense of the variogram estimates, particularly in the presence of noise. In this case it confirms the pattern in the variogram, reflecting smaller variances at the smallest lag distances.
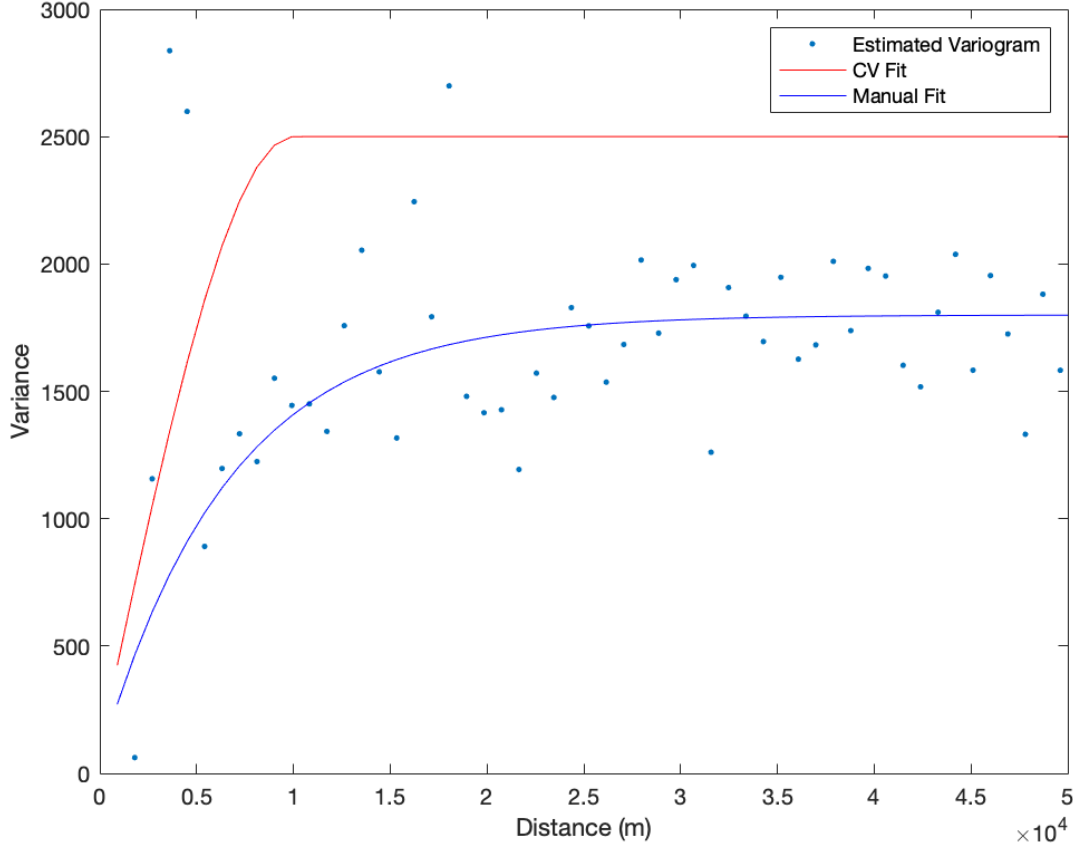

Figure 4: Estimated variogram plot from training data, with CV fit variogram and manual fit variogram overlaid.

Using the estimated variogram plot as reference I manually fit a variogram model to the estimated variogram points. The two variogram models that I tested were a spherical model defined as:

$$\hat{\gamma}(h) = \begin{cases} s^2(\frac{2h}{2r} - \frac{h^3}{2r^3} : h \leq r) \\ s^2 : otherwise \end{cases} \tag{5}$$

And an exponential model defined as:

$$\hat{\gamma}(h) = \begin{cases} a + ((s-a)(1 - e^{-3h/r})) : h > 0 \\ 0 : h == 0 \end{cases} \tag{6}$$

where h is the the lag distance between points, a is the nugget, r is the range, and s is the sill.

The parameters chosen for the manually fit variogram model can be seen in Table 2, and the fit is
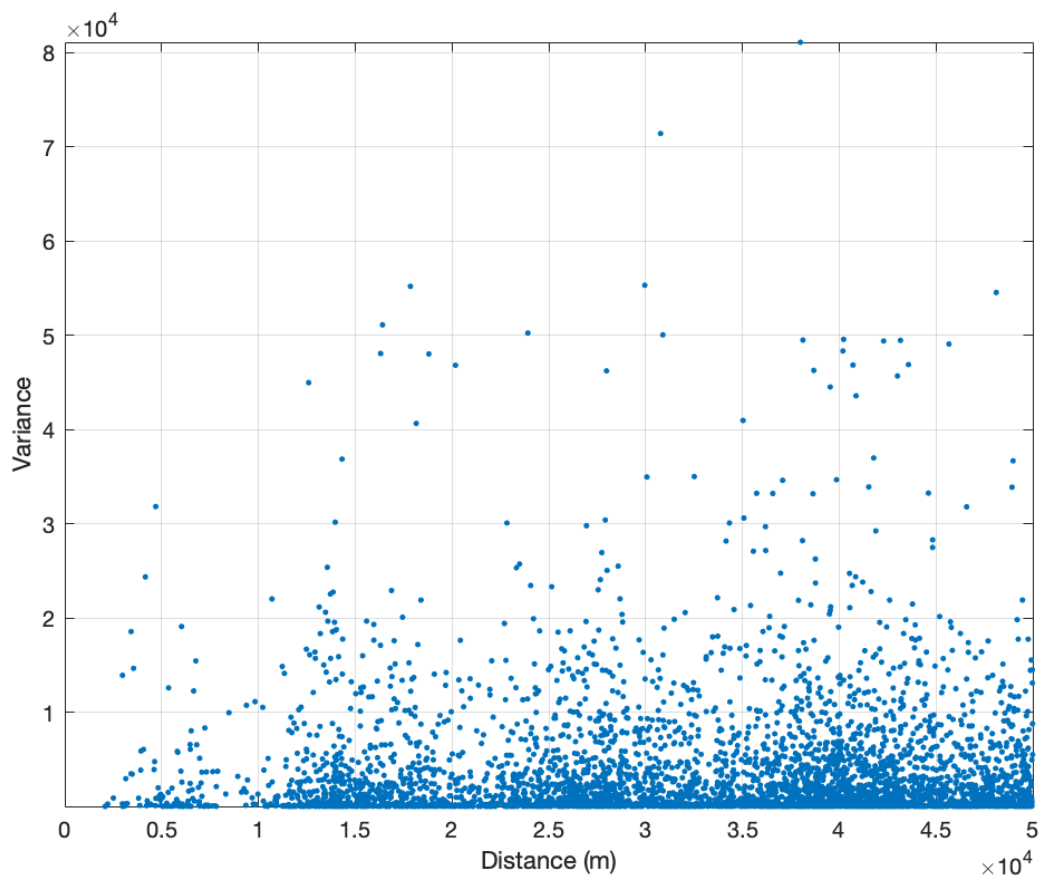
9

Figure 5: Variogram cloud from training data.

Table 2: Variogram parameters selected through CV fitting and manual fitting; Range in km.

Fit	Model Type	Sill	Range	Nugget
CV	Exponential	2,500	10	100
Manual	Exponential	1,800	20	50

exhibited by the blue curve in Figure 4. Additionally, I performed a standard grid search across the four variogram parameters (sill, range, nugget, model type) and assessed each using 10-fold cross-validation. I performed an assessment for each fold using ordinary kriging (described subsequently), and I assessed the performance by computing the root mean squared error (RMSE; calculation described subsequently) of the kriged predictions of residuals compared against the original model residuals. For each unique combination of variogram parameters, I randomly split the training data into 10 folds. For the kth iteration of the cross validation, the kth fold of the data was held out, and the k-1 other folds were used to produce the covariance matrix which is used to make kriging predictions. Predictions were made for the kth fold and those predictions were compared to the actual values. The RMSEs produced for each of the 10 iterations was averaged to produce a single RMSE value for each unique set of variogram parameters. I used the plotted variogram estimates (Figure 4) to determine suitable search ranges for each of the variogram parameters which are described in Table 3. These combinations were sorted by their associated RMSE values, and the combination with the lowest RMSE was selected. This best combination of parameters is described in Table 2, and the fit is shown by the red line in Figure 4. The CV RMSE produced by the various parameter combinations ranged from a low of of 39.8 AGB Mg ha^{-1} (selected iteration), to a high of 47.05 AGB Mg ha^{-}1.

Interestingly, the best combination of parameters selected by the CV-fit process produced a much higher sill than what I chose during my manual fit process. The CV fit approach seemed to track along some of the larger variogram estimates at roughly 4-5 km lag distances which I interpreted as outliers during my manual fitting process. In looking at the variogram cloud in Figure 6, which is just a re-plotting of Figures 4 and 5 with the x-axis restricted to 20km, it does appear that there are a few high-variance outliers at roughly 4-5km lag distances that are influencing the variogram estimates at these lags. Given the low number of pairs at these lags, these outliers will have a large impact on the variogram estimate at these ranges. Based on this analysis, I remain confident in my manually fit variogram - given the information available to me as an invested practitioner, I believe the chosen variogram parameters were sensible.

I used ordinary kriging to produce prediction surfaces across the state, and to produce point predictions at the testing data locations. Kriging estimates for a location $s0$ were computed as follows:

$$\hat{Y}(s_0) = \sum_{i=1}^{n} w_i Y(s_i) \tag{7}$$

Where $Y(s_i)$ is the value for sampled (member of training data set) location i and w_i is a corresponding weight computed as follows:

$$w_+(s) = C_+^{-1} c_+(s) \tag{8}$$

Table 3: Variogram parameter search ranges (Min to Max by Step); 2,112 unique combinations tested; model type 0 corresponds to the spherical model and model type 1 corresponds to the exponential model; Range in km.

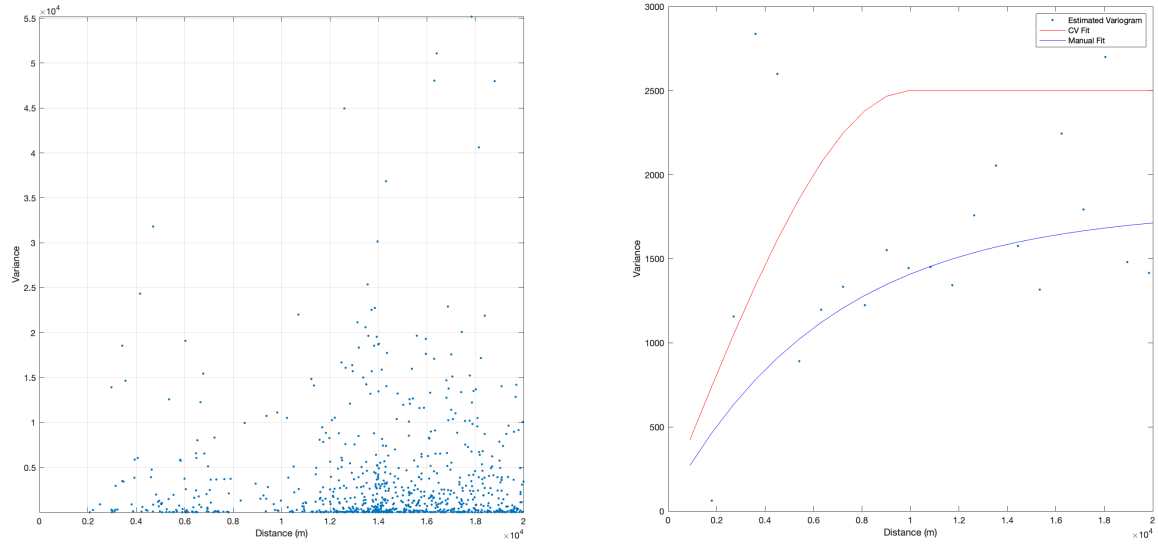| Parameter	Min	Max	Step
Model Type	0	1	1
Nugget	0	1,000	100
Range	10	25	1
Sill	1,250	2,500	250

Figure 6: Variogram cloud and variogram fitting restricted to 20km lag distances.

Where C_+^{-1} is the inverse covariance matrix computed with the provided variogram model and the sampled locations with an additional column and row of ones appended. The last value in the diagonal (bottom right) is set to 0. These modifications enforce that weights w sum to 1. c_+ is a vector containing covariance between the estimated location s and the sampled points, with a 1 appended to enforce weights w sum to 1. The $+$ subscript for w, represents the presence of the "Lagrange Multiplier" in the w vector, which is not used practically but is used in computation to force the sum of the weights to equal 1. The variance of each kriging prediction is computed as follows:

$$\sigma_e^2 = \sigma^2 - c_+^T(s)C_+^{-1}c_+(s) \tag{9}$$

Where σ^2 is the sill parameter.

The kriging surfaces produced with both variograms is displayed in Figure 7 and the associated variance surfaces are displayed in Figure 8. The kriging predictions from both variograms seem to correctly capture the range of residuals as present in the training data. Notably, the manually fit variogram predictions were less conservative, in that there were more predictions further from 0, as evidenced by the less proportion of white space in the manual variogram kriging prediction map. This is likely a function of the larger range parameter chosen for the manually fit variogram (Table 2). The variance surface (Figure 8) shows how quickly the uncertainty of the kriged predictions grows as distances from plot locations increase, as we can assume the dots or pockets of lower variance are plot locations. Displaying a similar pattern to the prediction surface, the variance of the CV fit variogram rises much faster as distances from the inferred plot locations increase. In both cases, we can say that the predictions were not very accurate with any significant distance from plot locations, but this plays out more strongly with the CV fit variogram predictions. Finally, the maximum variance produced by the CV fit variogram predictions was much larger than the variance of the manually fit variogram predictions due to the differences in the selected sill parameters.

Both the CV fit variogram and the manually fit variogram were used to make predictions at the testing data locations (displayed by the black points in Figures 7, 8). The residuals predicted at these locations were subtracted from the original model predictions at these locations to create new "improved" estimates. The new estimates, as well as the original estimates, were compared to the FIA AGB values. RMSE (Equation (10)), mean bias error (MBE; Equation (11)), and the coefficient of determination (R^2; Equation (12)) were computed for the three sets of predictions (original, CV fit predictions, and manual predictions). These three performance metrics were computed as follows:

$$\text{RMSE} = \sqrt{(\frac{1}{n}) \sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{10}$$

$$\text{MBE} = (\frac{1}{n}) \sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{11}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{12}$$

Where n is the number of FIA plots included in the data set, $\hat{y}_i$ is the predicted value, y_i the
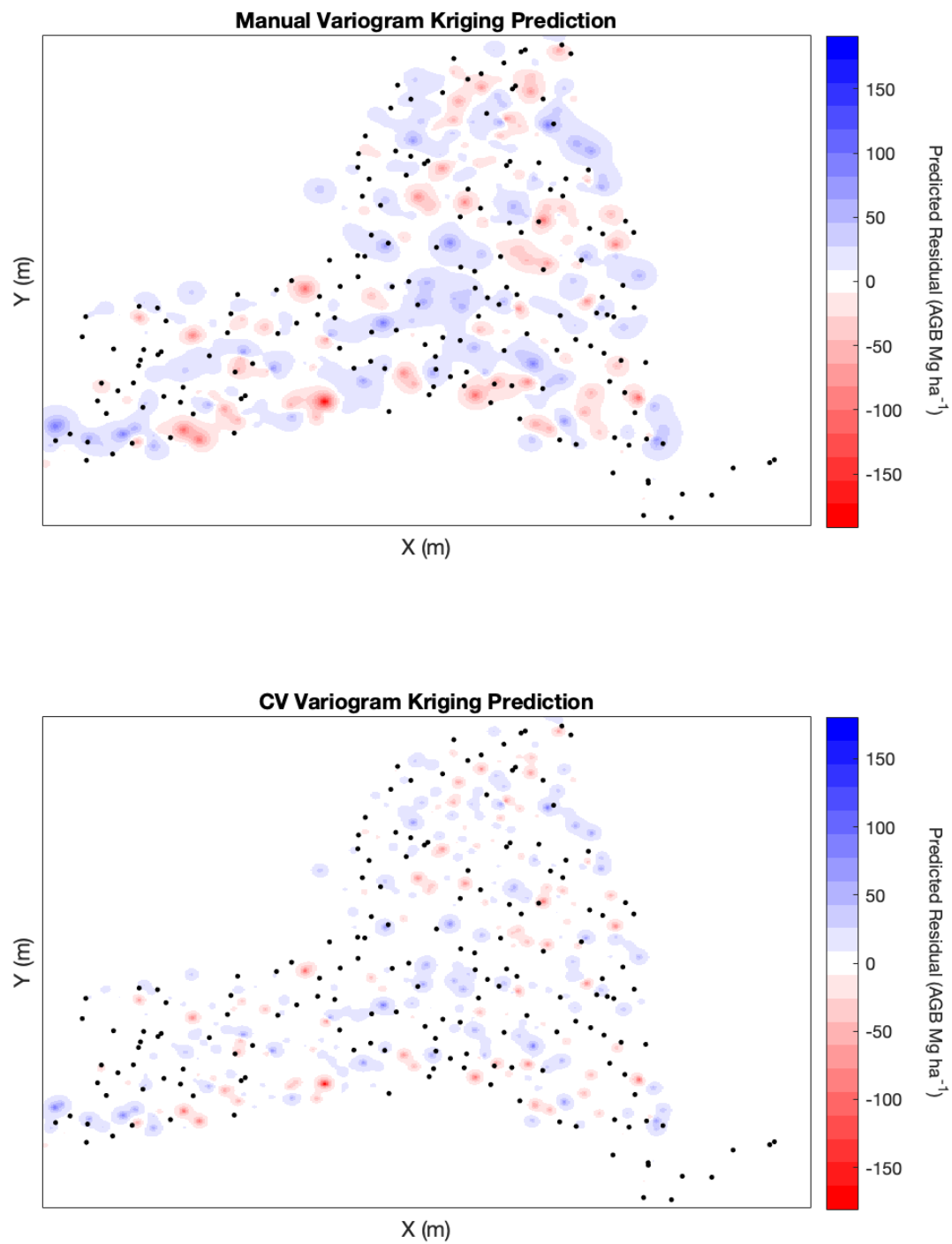
13

Figure 7: Kriging prediction surfaces for two variogram fitting approaches with test plot locations overlaid. X and Y axis marks removed to preserve the confidentiality of the FIA plot locations.
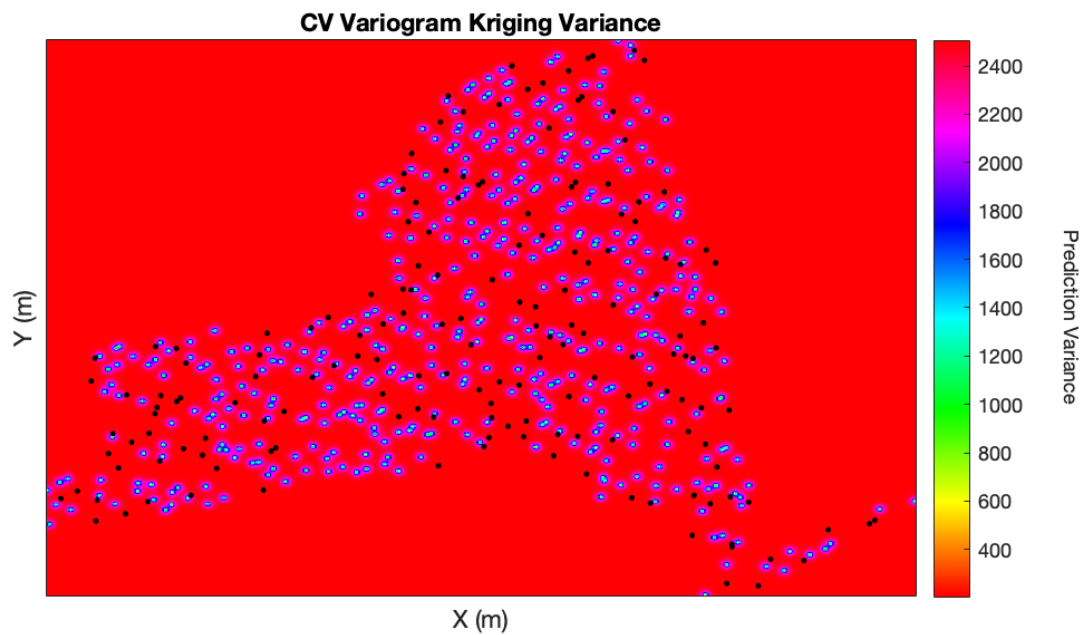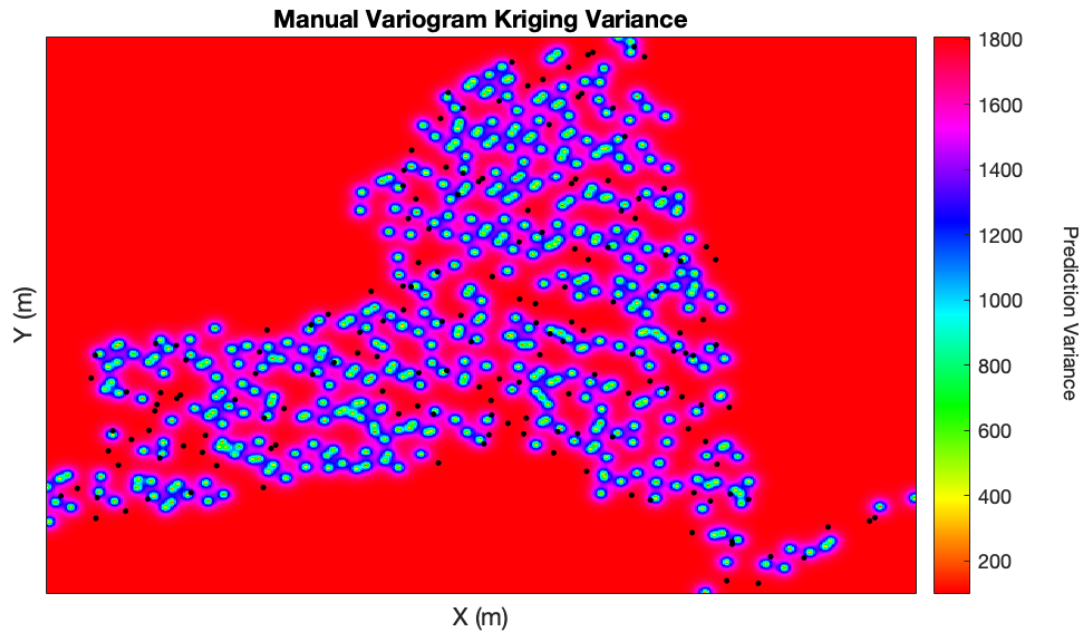
Figure 8: Kriging variance surface for two variogram fitting approaches with test plot locations overlaid. X and Y axis marks removed to preserve the confidentiality of the FIA plot locations.

Table 4: Performance metrics for model accuracy at testing locations before and after prediction improvement via residual kriging; RMSE, MBE in AGB Mg ha^{-1}.

Predictions	RMSE	MBE	R^2
Original	42.69	-2.24	0.76
CV Fit	42.61	-5.82	0.76
Manual Fit	43.39	-5.56	0.75

measured value at the corresponding location, and $\bar{y}$ the mean value from measurements or observations.

Differences between all three sets of predictions were marginal, however the CV fit kriging predictions were the best in terms of RMSE, while the original (unimproved by kriging) predictions were distinctly better than both kriging approaches in terms of MBE (Table 4). Interestingly, the CV fit approach had marginally worse MBE compared to the manually fit predictions, despite producing a better RMSE. Additionally, each of the three sets of predictions were plotted against the FIA reference values and a 1 to 1 line was overlaid (Figure 9). In most cases, all three predictions were very similar which further reinforces the comparison of accuracy metrics in Table 4. The general pattern across all three sets of predictions was overprediction on the low end of the reference AGB distribution and underprediction on the high end of reference AGB distribution. In other words the predictions were biased towards the mean. This is undoubtedly a function of the model itself rather than the spatial interpolation process given the similarity of the kriging predictions to the original model predictions. The increased size of the bias (more negative) in the kriging improved predictions might be better understood by the kriging prediction surfaces which were dominated by positive residuals (blue in Figure 7). Positive predicted residuals would result in decreased predictions after subtracting the residuals from the original predictions. It is possible that the abundance of overpredictions on reference locations with 0 AGB (see the Y-Axis of Figure 9) in the original model predictions led to the overabundance of positive residuals in the predicted surface and subsequently an overabundance of positive residuals at the testing locations.

Overall, regression kriging with both variogram fitting approaches did little to improve the model predictions at the testing locations. This is likely due to weak spatial relationship between residuals at FIA plot locations as evidenced by the small range in variance in the variogram estimate plot (Figure 4). Perhaps more importantly, the range of the variograms, or rather the distances at which information could be gathered from other locations was quite short (Table 2), compared to the average distance between plots (Table 1). As evidenced by the prediction surface (Figure 7), very conservative (near-zero) predictions were made with any significant distance from the original training plots, and it was rare that the testing plot locations fell close enough to a training plot for the prediction there to be significantly improved.

Conclusions

This report showed that regression kriging did little to improve Landsat-based AGB model predictions at FIA plots inventoried in 2019 throughout NYS. The spatial relationships of the model residuals were not generalizable beyond very short distances, and the spacing of FIA plots
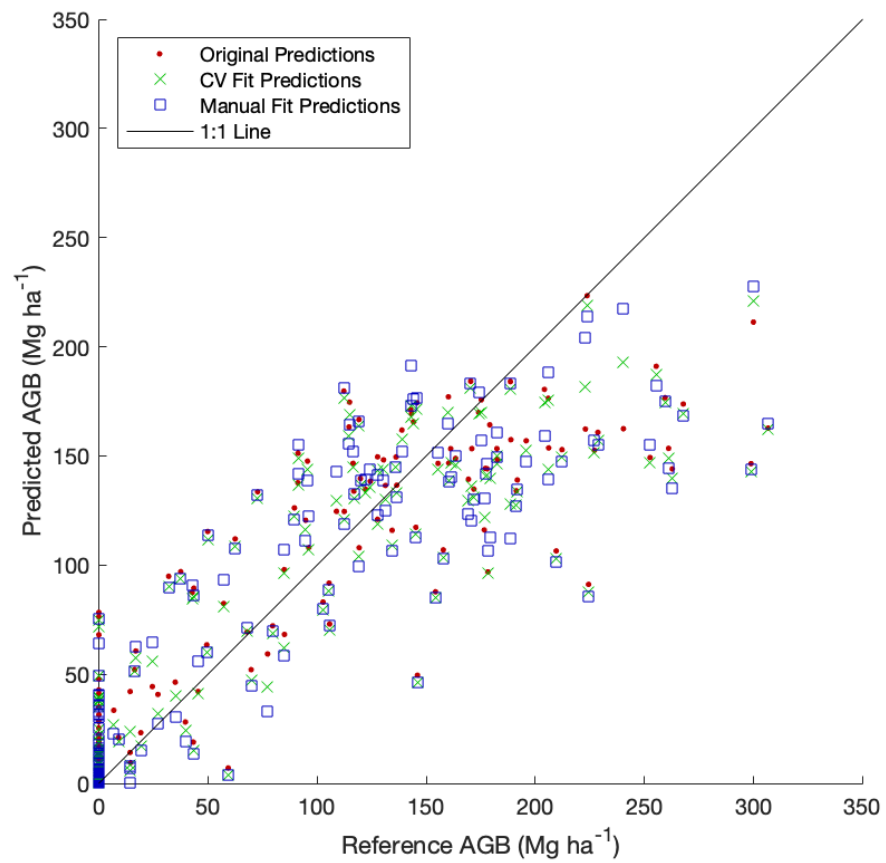
Figure 9: Predicted vs reference scatter plot for three sets of predictions at test data locations.

does not support the transfer of information from plot to plot. Additionally, we might conclude that the model residuals are mostly randomly distributed in space, indicating that the models are properly calibrated. While this is good news for the models themselves, it certainly does not help the kriging process. For this approach to be more effective we might require a more intensive sampling protocol, or rather a random sample instead of a systematic one to create a wider range of plot-to-plot distances that could be leveraged for spatial interpolation. I was able to demonstrate that a grid search with k-fold cross-validation produced a better variogram fit than what a human practitioner invested in the outcome could produce, as evidenced by the marginally more accurate kriging predictions for the CV fit variogram as compared to those made with the manually fit variogram.

Future Work

This approach might benefit from breaking down the study area into smaller, more homogeneous, units to promote covariance between plots. Eco-regions might be suitable for this approach, however care would need to be taken to ensure that a sufficient number of observations were present in each unit. Additionally, this work could be extended by including all five annual sampling frames to include the full inventory of NYS plots. Plots inventoried from 2019 were selected in this study for efficiency gains and temporal consistency in the results, but the kriging process might benefit from a more dense set of observations. Finally, this approach might be more successful if individual FIA subplots were treated as individual units, thus allowing for more fine-scale spatial relationships in the data. From the outset of my modeling approach I treated the aggregate of the 4 subplots as a single unit, trading smaller sample size for more information at each unit and a reduction in plot-level AGB variance. Given this approach, a model residual produced at the subplot was incongruous with our model design.

References

Banskota, Asim, Nilam Kayastha, Michael J. Falkowski, Michael A. Wulder, Robert E. Froese, and Joanne C. White. 2014. "Forest Monitoring Using Landsat Time Series Data: A Review." *Canadian Journal of Remote Sensing* 40 (5): 362–84. https://doi.org/10.1080/07038992.2014. 987376.

Bechtold, William A, and Paul L Patterson. 2005. *The Enhanced Forest Inventory and Analysis Program–National Sampling Design and Estimation Procedures*. Vol. 80. USDA Forest Service, Southern Research Station.

Brown, Jesslyn F., Heather J. Tollerud, Christopher P. Barber, Qiang Zhou, John L. Dwyer, James E. Vogelmann, Thomas R. Loveland, et al. 2020. "Lessons Learned Implementing an Operational Continuous United States National Land Change Monitoring Capability: The Land Change Monitoring, Assessment, and Projection (LCMAP) Approach." *Remote Sensing of Environment* 238: 111356. https://doi.org/10.1016/j.rse.2019.111356.

Chen, Qi, and Ronald McRoberts. 2016. "Statewide Mapping and Estimation of Vegetation Aboveground Biomass Using Airborne Lidar." In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE. https://doi.org/10.1109/igarss.2016.7730157.

Daly, Christopher, Michael Halbleib, Joseph I. Smith, Wayne P. Gibson, Matthew K. Doggett, George H. Taylor, Jan Curtis, and Phillip P. Pasteris. 2008. "Physiographically Sensitive Map-

ping of Climatological Temperature and Precipitation Across the Conterminous United States." *International Journal of Climatology* 28 (15): 2031–64. https://doi.org/10.1002/joc.1688.

DEC. n.d. "Forests." *Forests - NYS Dept. Of Environmental Conservation*. NYS Dept. of Environmental Conservation. https://www.dec.ny.gov/lands/309.html.

Dwyer, John L., David P. Roy, Brian Sauer, Calli B. Jenkerson, Hankui K. Zhang, and Leo Lymburner. 2018. "Analysis Ready Data: Enabling Analysis of the Landsat Archive." *Remote Sensing* 10 (9). https://doi.org/10.3390/rs10091363.

Freeman, Elizabeth A., and Gretchen G. Moisen. 2006. "Evaluating Kriging as a Tool to Improve Moderate Resolution Maps of Forest Biomass." *Environmental Monitoring and Assessment* 128 (1-3): 395–410. https://doi.org/10.1007/s10661-006-9322-6.

Gray, Andrew N, Thomas J Brandeis, John D Shaw, William H McWilliams, and Patrick Miles. 2012. "Forest Inventory and Analysis Database of the United States of America (FIA)." *Biodiversity and Ecology* 4: 225–31. https://doi.org/10.7809/b-e.00079.

Homer, Collin, and A Gallant. 2001. "Partitioning the Conterminous United States into Mapping Zones for Landsat TM Land Cover Mapping." *Unpublished US Geologic Survey Report. Http://Landcover. Usgs. Gov/Pdf/Homer. Pdf.(1 August 2008)*.

Huang, Wenli, Katelyn Dolan, Anu Swatantran, Kristofer Johnson, Hao Tang, Jarlath O'Neil-Dunne, Ralph Dubayah, and George Hurtt. 2019. "High-Resolution Mapping of Aboveground Biomass for Forest Carbon Monitoring System in the Tri-State Region of Maryland, Pennsylvania and Delaware, USA." *Environmental Research Letters* 14 (9): 095002. https://doi.org/10.1088/1748-9326/ab2917.

Hudak, Andrew T, Patrick A Fekety, Van R Kane, Robert E Kennedy, Steven K Filippelli, Michael J Falkowski, Wade T Tinkham, et al. 2020. "A Carbon Monitoring System for Mapping Regional, Annual Aboveground Biomass Across the Northwestern USA." *Environmental Research Letters* 15 (9): 095003. https://doi.org/10.1088/1748-9326/ab93f9.

Hudak, AT, Michael Lefsky, Warren Cohen, and Mercedes Berterretche. 2002. "Integration of Lidar and Landsat ETM+ Data for Estimating and Mapping Forest Canopy Height." *Remote Sensing of Environment* 82 (October). https://doi.org/10.1016/S0034-4257(02)00056-1.

Hurtt, G, M Zhao, R Sahajpal, A Armstrong, R Birdsey, E Campbell, K Dolan, et al. 2019. "Beyond MRV: High-Resolution Forest Carbon Modeling for Climate Mitigation Planning over Maryland, USA." *Environmental Research Letters* 14 (4): 045013. https://doi.org/10.1088/1748-9326/ab0bbe.

Karatzoglou, Alexandros, Alexandros Smola, Kurt Hornik, and Achim Zeileis. 2004. "Kernlab - an S4 Package for Kernel Methods in r." *Journal of Statistical Software, Articles* 11 (9): 1–20. https://doi.org/10.18637/jss.v011.i09.

Ke, Guolin, Damien Soukhavong, James Lamb, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2021. *Lightgbm: Light Gradient Boosting Machine*. https://CRAN.R-project.org/package=lightgbm.

Kennedy, Robert E, Janet Ohmann, Matt Gregory, Heather Roberts, Zhiqiang Yang, David M Bell, Van Kane, et al. 2018. "An Empirical, Integrated Forest Biomass Monitoring System." *Environmental Research Letters* 13 (2): 025004. https://doi.org/10.1088/1748-9326/aa9d9e.

Mahoney, Michael. 2021. *terrainr: Landscape Visualizations in r and Unity*. https://doi.org/https://doi.org/10.5281/zenodo.5142763.

MATLAB. 2021. *Version 9.11.0.1769968 (R2021b)*. Natick, Massachusetts: The MathWorks Inc.

Meng, Qingmin, Chris Cieszewski, and Marguerite Madden. 2009. "Large Area Forest Inventory

Using Landsat ETM+: A Geostatistical Approach." *ISPRS Journal of Photogrammetry and Remote Sensing* 64 (1): 27–36. https://doi.org/https://doi.org/10.1016/j.isprsjprs.2008.06.006.

Riemann, Rachel, Barry Tyler Wilson, Andrew Lister, and Sarah Parks. 2010. "An Effective Assessment Protocol for Continuous Geospatial Datasets of Forest Characteristics Using USFS Forest Inventory and Analysis (FIA) Data." *Remote Sensing of Environment* 114 (10): 2337–52. https://doi.org/10.1016/j.rse.2010.05.010.

Simard, Marc, Naiara Pinto, Joshua B Fisher, and Alessandro Baccini. 2011. "Mapping Forest Canopy Height Globally with Spaceborne Lidar." *Journal of Geophysical Research: Biogeosciences* 116 (G4).

Tsui, Olivier W., Nicholas C. Coops, Michael A. Wulder, and Peter L. Marshall. 2013. "Integrating Airborne LiDAR and Space-Borne Radar via Multivariate Kriging to Estimate Above-Ground Biomass." *Remote Sensing of Environment* 139: 340–52. https://doi.org/https://doi.org/10.1016/j.rse.2013.08.012.

Wolpert, David H. 1992. "Stacked Generalization." *Neural Networks* 5 (2): 241–59. https://doi.org/10.1016/S0893-6080(05)80023-1.

Wright, Marvin N., and Andreas Ziegler. 2017. "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in c++ and r." *Journal of Statistical Software, Articles* 77 (1): 1–17. https://doi.org/10.18637/jss.v077.i01.

Wulder, Michael A., Jeffrey G. Masek, Warren B. Cohen, Thomas R. Loveland, and Curtis E. Woodcock. 2012. "Opening the Archive: How Free Data Has Enabled the Science and Monitoring Promise of Landsat." *Remote Sensing of Environment* 122: 2–10. https://doi.org/https://doi.org/10.1016/j.rse.2012.01.010.

Zhu, Zhe, and Curtis E. Woodcock. 2014. "Continuous Change Detection and Classification of Land Cover Using All Available Landsat Data." *Remote Sensing of Environment* 144: 152–71. https://doi.org/10.1016/j.rse.2014.01.011.