TESTING THE IMPACT OF USING AUXILIARY VARIABLES IN THE DEFINITION OF *K*-NEAREST NEIGHBOR IMPUTATION

by

MANLIN WEI

(Under the Direction of CHRIS CIESZEWSKI)

ABSTRACT

In this study, we use the *k*-Nearest Neighbor (*k*NN) method to test the potential of computing Georgia forest volumes based from the FIA plot data and Landsat ARD with auxiliary variables such as maximum and minimum temperature, land cover, and elevation. The total number of plots used is 941, we divided them into the training dataset (708) and testing dataset (233). The training dataset was used to calibrate the *k*NN model, and the testing dataset was used to test the model. The results showed that the best *k*NN model was a combination of ARD and all auxiliary variables, the RMSE (%) and bias are 77.61% and -5.40%, respectively. Using the derived results, we have created a volume map, used subsequently for the county-level volume estimates.

INDEX WORDS: *k*-nearest neighbor, Landsat ARD, auxiliary variables, forest volume

TESTING THE IMPACT OF USING AUXILIARY VARIABLES IN THE DEFINITION OF

K-NEAREST NEIGHBOR IMPUTATION

by

MANLIN WEI

BS, Beijing University of Agriculture, China, 2018

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2020

TESTING THE IMPACT OF USING AUXILIARY VARIABLES IN THE DEFINITION OF *K*-NEAREST NEIGHBOR IMPUTATION

by

MANLIN WEI

Major Professor:	Chris Cieszewski
Committee:	Pete Bettinger
	Roger Charles Lowe

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
August 2020

DEDICATION

To myself

We are loved by the universe.

Be patient, step by step.

ACKNOWLEDGEMENTS

Thanks to my mom who gives me support to finish this degree. Thanks to my major professor, Dr. Cieszewski, who let me do this research and advised me during my study. Thanks to Dr. Bettinger who gave me the opportunity as the FANR 3000 teaching assistant in 2019 and advised me to attend conferences. I learned a lot of things during the time as a TA and felt inspired at the conferences. Thanks to Dr. Lowe who taught me how to use ArcGIS, that was the first class I took in the U.S.A. Thanks to Shingo Obata, a great guy who offered me a lot of help in my research. And thanks to Taeyoon Lee, Rocio Gutierrez, Volkan Bektas, for sharing a lot of meaningful time during my study.

Thanks to my friend: Shiyi Wang. Even though you are back in China, I would like to tell you the time we were supporting each other at UGA was valuable. I hope both of us will have a bright future.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Background

FIA Inventory

The United States Forest Service (USFS) was created in 1905 to protect federally-owned natural resources across the United States. Nowadays, the agency mission is to “To sustain the health, diversity, and productivity of the Nation’s forests and grasslands to meet the needs of present and future generations” (United States Forest Service, 2012). The focus is on ecosystem health, and it represents a combination of past acts and policy. However, the USFS is an organization much broader than simply the managers of national forests and grasslands. The USFS also has a research program that is separate from the National Forest System. Within this research program are many scientists who explore all aspects of forests and lands and the resources that can be found there. Further, the USFS employs researchers who seek to expand the use of remote sensing technology to monitor and describe forest resources.

The Forest Inventory and Analysis (FIA) program of the USFS has been operating since 1930 to collect and report the nation’s forest data. USFS field crews conduct forest inventories on lands belonging to private owners, federal, state, and local governments. Along with the changes in the USFS mission, the FIA inventory has evolved, and further, measurement activities have changed from periodic to annual, and state inventory is reported every five years. Every year, 20% of the total plots are randomly selected and measured in most states. The sampling design of FIA plots has three phases that represent the structure of the program (Bechtold and Patterson, 2005).

1. Phase 1 is to stratify the land area to improve the accuracy of the estimates.
2. Phase 2 is to visit and measure the traditional FIA suite of variables.
3. Phase 3 is to visit and measure additional variables related to forest health conditions and ecosystems.

The data we use is the inventory data of Georgia in 2017 measured by the FIA. In 2017, the FIA reported that the forestland of Georgia was 9.9 million hectares with a 0.52% sampling error. The average annual net growth of live trees (at least 0.1 meters d.b.h/d.r.c) was 57.9 million cubic meters per year with a 1.52% sampling error. The total number of inventory plots in Georgia are 4,903, and 941 plots were measured in 2017. The major tree species in Georgia are loblolly pine *(Pinus taeda)*, slash pine *(Pinus elliottii)*, sweetgum *(Liquidambar styraciflua)*, swamp tupelo *(Nyssa biflora)*, yellow-poplar *(Liriodendron tulipifera)*, water oak *(Quercus nigra)*, red maple *(Acer rubrum)*, longleaf pine *(Pinus palustris)*, and cypress *(Taxodium spp.)* (Lambert 2020).

Landsat Imagery

Landsat imagery is the product of the Landsat mission, Landsat 1-9 were different satellites used to observe the earth (National Aeronautics and Space Administration, 2020). The USFS and many research organizations use Landsat imagery to describe various phenomena on Earth.

1. Landsat 1 was operated from 1972-1978 as the first Earth-observing satellite. The instruments in Landsat 1 included a camera system called the return beam vidicon (RBV) and the multispectral scanner (MSS).

2. Landsat 2 was operated from 1975-1982. It was launched three years after Landsat 1 with the same instruments as Landsat 1.
3. Landsat 3 was operated from 1978-1983. It was launched with the same instruments as predecessors but RBV had a small improvisation.
4. Landsat 4 was operated from 1982-1993 without carrying RBV. Besides MSS, it also carried a Thematic Mapper (TM) sensor.
5. Landsat 5 was operated from 1994-2013 with the same instruments as Landsat 4. However, in 1995 the MSS was turned off due to technical issues of Ku-band transmitter.
6. Landsat 6 was launched in 1993 but failed because not reaching enough velocity to orbit the Earth.
7. Landsat 7 was operated from 1999-current with the Enhanced Thematic Mapper Plus (ETM+) which was an improvement over the TM sensor.
8. Landsat 8 has been operating since 2013 with the Operational Land Imager (OLI) and the Thermal Infrared Sensor (TIRS).
9. Landsat 9 will be launched in 2021.

Users can download Landsat images using the Earth Explorer website (https://earthexplorer.usgs.gov/) free of charge. The accessibility and availability of Landsat images allow researchers to utilize this resource. However, the pre-processing steps for Landsat imagery can be complex (Dwyer et al., 2018). These steps are necessary because they ensure the Landsat data is correctly displayed in time and space. Landsat ARD was released by the United States Geological Survey (USGS) in 2017, so it is a relatively new data resource based on Landsat Collection 1 Level-1 input. The purpose of releasing ARD is to reduce the burden of pre-processing.

Though ARD can be derived from many Landsat products, the data we are using are the ARD images only from Landsat 8 mission. Landsat 8 has 7 bands, each band represents the wavelength of electromagnetic energy, reflected or emitted from the earth. The pixel value stored in the band indicates the amount of energy the satellite could capture and stored in a specific wavelength (Table 1).

Table 1. The bands in Landsat 8

Band number	Radiation	Wavelength (*nm*)
1	Deep blue and violets	433-453
2	Visible blue	450-515
3	Visible green	525-600
4	Visible red	630-680
5	Near infrared	845-885
6	Shortwave infrared	1560-1660
7	Shortwave infrared	2100-2300

Literature Review of the Nearest Neighbor Method in Forestry

The nearest neighbor (NN) method is often used to locate the closest unit from other units in a feature space. Units in the space are assumed to have similar characteristics. Then, the closest unit is determined by a vector distance between two units. Often, the NN method is used to classify some unknown elements or fill missing data in a dataset. When there are large amounts of unknown or missing data, this approach becomes powerful as unlike the linear model, the NN method does not need to make any assumptions about the distribution of data and non-parametric is needed in this method. Therefore, we can use this approach without fully defining the relationship between the predicted elements and the observed elements. In general, more than one nearby unit is used to determine the nearest neighbor. Therefore, a generalization term of the nearest neighbor method is called the *k*-Nearest-Neighbor method as we will find *k* closest units to determine the nearest neighbor.

Some observational data, called X-variables need to be used to define the feature space in NN imputation. Remote sensing data such as lidar, satellite imagery, aerial photography is often used as the X-variables. The first well-known application in forestry was a study in Finland (Tomppo and Katila, 1991), this study accomplished a systematic *k*NN imputation to forest variables in the country. Then, more and more attention were drawn to this method. Some studies used *k*NN to estimate forest attributes such as basal area (LeMay and Temesgen, 2005; Meng et al., 2007), tree canopy density (Yazdani et al., 2020), volume (Tompalski et al., 2019), and biomass (Nguyen et al., 2018). Other studies used *k*NN for classification purposes, for example, for classifying forestland using Landsat imagery (McRoberts et al., 2002), classify the landcover type (Ge et al., 2020). Some studies have also demonstrated using the *k*NN method with spatial data to map forest attributes (Franco-Lopez et al., 2001; Trotter et al., 1997).

Some issues were also discussed using NN methods. The relative RMSE of NN imputation is often high, and one paper mentioned that most studies the RMSE were around 60% (Tanaka et al., 2015). NN imputation is also inherently biased because the predicted volume will be no larger than the largest response variable or smaller than the smallest response variable (McRoberts, 2009). Further, one study found there is an association between the residuals and the observed values, indicating a high variation of observations will increase the residuals (Kajisa et al., 2008).

Purpose of the study

In forestry, we usually cannot measure all trees in a study area because the total amount of work will cost so much time and labor. Instead, we usually set up some sample plots in the forest, assuming the sample plots can represent essential inventory information in the whole study area. Though there are also other sampling methods, the core idea is to use some samples to represent the whole population. Therefore, it is common that the forest inventory is not a complete dataset as we only measured sample plots instead of all trees in the forest. Since the FIA program measures 20% of plots per year, it will take 5 years until the inventory is completed to get yearly total inventory within this circle. What if we want to have up-to-date inventory data? The nearest neighbor imputation method we introduced previously can be an option to answer this question.

Tomppo and Katila (1991) used *k*NN to get up-to-date inventory information for an area smaller than 150, 000 ha which is the smallest area ground systematic cluster sampling that they could compute reliably. While some studies using *k*NN imputation have been based on satellite imagery (Kajisa et al., 2008; Meng et al., 2007; Gjertsen, 2007; Gu et al., 2006), only a few studies using auxiliary variables combined with satellite imagery (Tomppo and Katila, 1991; Tanaka et al., 2015). To understand whether auxiliary variables can improve the accuracy of *k*NN imputation, we will combine satellite imagery with auxiliary variables to do *k*NN imputation.

The objective of the study is to use the *k*-Nearest-Neighbor method to test the potential of computing forest volume for the State of Georgia using Landsat imagery with auxiliary variables based on FIA data.

Expected Results

We expect the accuracy of *k*NN imputation using auxiliary variables will lead to better results with less error and bias than when using *k*NN imputation without auxiliary variables.

CHAPTER 2

TESTING THE IMPACT OF USING AUXILIARY VARIABLES IN THE DEFINITION OF *K*-NEAREST NEIGHBOR IMPUTATION

Introduction

The *k*-Nearest-Neighbor is a non-parametric method that is capable of filling missing data using some known data in the dataset. The missing data is calculated by its nearest or similar neighbor in the feature space. Feature space can be understood as a vector space of variables. Defining the nearest neighbor in the feature space is an essential part of doing *k*NN imputation as this will directly affect the accuracy of prediction. Therefore, many studies explored various methods to calculate distance metrics (McRoberts, 2012; LeMay and Temesgen, 2005). Studies applied different distance metrics include Euclidean distance (LeMay and Temesgen, 2005), Mahalanobis distance (Stage and Crookston, 2007), most similar neighbor (MSN) (Moeur and Stage, 1995), gradient nearest neighbor (GNN) (Ohmann and Gregory, 2002). Selecting distance metrics is based on the objective of the study. Determining the value of *k* is also crucial for *k*NN imputation. Small *k* values usually have large RMSE but maintain the variation of the prediction, while large *k* values have small RMSE with less variation of the prediction (Eskelson et al., 2009). The most commonly used method for choosing *k* is selecting the smallest *k* value in which RMSE is no more than 1% or 5% than the smallest RMSE which originated from a study (McRoberts et al., 2002). *k*NN can be used to predict categorical variables or continuous variables. The requirement for doing *k*NN is flexible, unlike linear or non-linear regression, we do not need to define the relationship between the predictions and the observations (Eskelson et al., 2009). However, we need to determine the value of *k* and the distance metrics before doing *k*NN. The *k*NN method has been applied in forestry for decades. In forestry, it has been widely used with forest inventory data. Some studies used *k*NN to estimate forest attributes such as basal area, tree canopy density (Yazdani et al., 2020), volume (Tompalski et al., 2019), and biomass (Nguyen et al., 2018). Other studies used *k*NN for classification purposes, for example, classifying forestland using Landsat imagery (McRoberts et al., 2002), classifying the landcover type (Ge et al., 2020).

There are many spatial datasets available for research purposes such as Lidar, Sentinel imagery, Landsat imagery, aerial photos, and so on. Sentinel imagery and Landsat imagery are satellite-derived images, they are time-series spatial data. Lidar is usually used for relatively small areas. Aerial photos can be collected for areas as large as states in the United States. Landsat imagery is the product of the Landsat mission, operated by The National Aeronautics and Space Administration (NASA). The spectral value in each pixel represents the amount of electromagnetic energy reflected or emitted from Earth, therefore it indicates the amount of energy the satellite could capture and stored in the image. The available Landsat imagery comes from Landsat 1 to Landsat 8, Landsat 9 will be launched in March 2021. Landsat ARD became available in 2017. The ARD dataset is based on Landsat Collection 1 Level-1 input. Landsat ARD does not need to be georeferenced to a known map projection. The Albers Equal Area Conic projection and WGS84 datum are used to represent the ARD. The Landsat ARD tiles use horizontal and vertical tile coordinates. The Landsat ARD includes Landsat 4-5 Thematic Mapper (TM), Landsat 7 Enhanced Thematic Mapper Plus (ETM+) and Landsat 8 Operational Land Imager (OIL)/Thermal infrared Sensor (TIR) imagery, which are provided with nonoverlapping tiles of 5000 × 5000 30 m pixels (Horizontal/Vertical Tile). Resampling of satellite imagery decreases the geometric fidelity of the resulting products due to pixel shifts or data smoothing (Dwyer et al., 2018). Landsat ARD is calibrated and re-projected directly into the ARD Albers Equal Area Conic projection, which means that ARD is resampled only once. The terrestrial variables in the ARD product such as Top of Atmosphere (TOA) reflectance, TOA Brightness Temperature (BT), Surface Reflectance (SR), and so on are projected directly into the uniform gridded system of ARD with metadata. Some previously published studies of using *k*NN imputation based on Landsat imagery (Tanaka et al., 2015; Tomppo et al., 2008; Meng et al., 2007), used older Landsat products. So we will use Landsat ARD as the spectral data in our study.

While many studies using *k*NN imputation have been based on satellite imagery (Tomppo, 1999; Tanaka et al., 2015; Kajisa et al., 2008; Meng et al., 2007; Gjertsen, 2007; Gu et al., 2006), only a few using auxiliary variables combined with satellite data (Tomppo, 1999; Tanaka et al., 2015). To understand whether auxiliary variables are efficient in *k*NN imputation, the objective of the study is to test the potential of computing forest volume of Georgia in 2017 using Landsat ARD with some auxiliary variables in *k*NN imputation.

Methods

Study Site

The study site is Georgia state of the United States. The total land area is 15.01 million hectares. Major tree species in Georgia are loblolly pine *(Pinus taeda)*, slash pine *(Pinus elliottii)*, sweetgum *(Liquidambar styraciflua)*, swamp tupelo *(Nyssa biflora)*, yellow-poplar *(Liriodendron tulipifera)*, water oak *(Quercus nigra)*, red maple *(Acer rubrum)*, longleaf pine

(Pinus palustris), and cypress *(Taxodium spp.)* (Lambert, 2020). The 12 Landsat ARD tiles cover entire Georgia were downloaded from Earth Explorer (https://earthexplorer.usgs.gov/) (Figure 1, Table 2). Though Landsat ARD can have many data sources, we only used the Landsat ARD products from Landsat 8 mission (Table 1). Of the terrestrial variables in the ARD product such as Top of Atmosphere (TOA) Reflectance, TOA Brightness Temperature (BT), Surface Reflectance (SR), we only used SR in this study. The spatial resolution or pixel size of ARD is 30 m. The cloud cover in all tiles was less than 10%.

Auxiliary Data

Some auxiliary data such as temperature and elevation are the environmental factors that affect tree growth. The other digital map, landcover can classify different land cover types to improve imputation accuracy (Tomppo and Katila 1991). The auxiliary data used in this study are minimum temperature, maximum temperature, land cover, and elevation downloaded from the Geospatial Data Gateway (United States Department of Agriculture, Natural Resources Conservation Service, 2019). Though the publication dates of some variables are before 2017 (Table 3), they are the latest published version we can download. Each variable contains only one band, each band will combine with Landsat ARD bands directly in the imputation process.

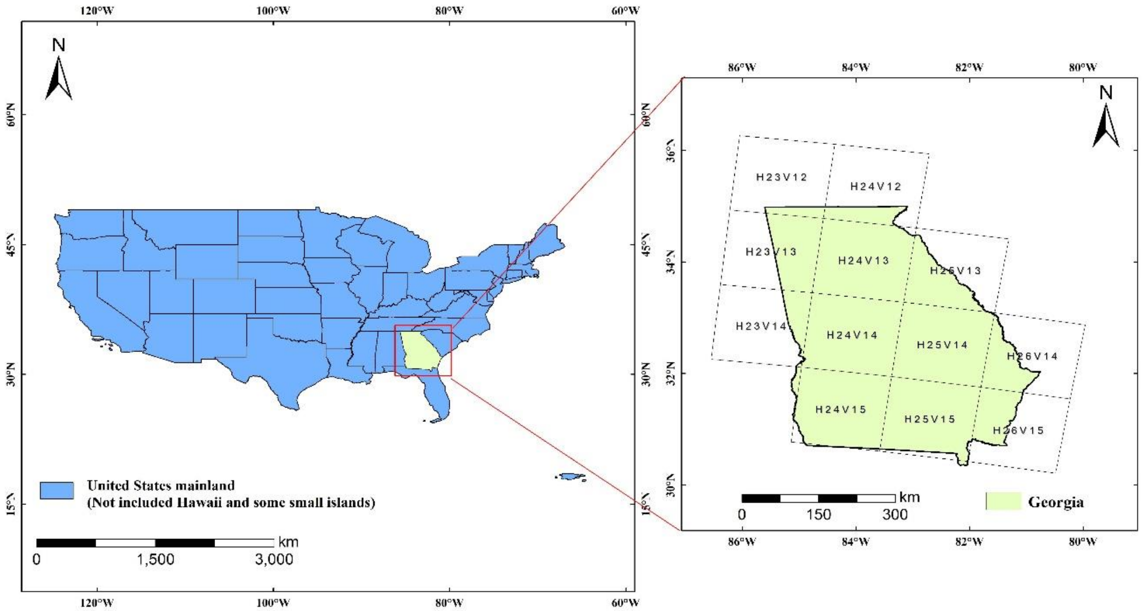

Figure 1. 12 Landsat ARD scenes cover the study location.

Table 2. The Landsat ARD tiles used in this study.

Landsat ARD tile	The number of Landsat scenes used in the tile	Date
H23V12	2	05/14/2017; 09/10/2017
H23V13	2	05/14/2017; 09/10/2017
H23V14	1	05/14/2017
H24V12	1	05/07/2017
H24V13	2	05/07/2017; 05/14/2017
H24V14	2	05/07/2017; 05/14/2017
H24V15	2	05/07/2017; 05/14/2017
H25V13	2	05/07/2017; 05/14/2017
H25V14	2	05/07/2017; 05/14/2017
H25V15	2	05/07/2017; 05/14/2017
H26V14	2	05/07/2017; 05/14/2017
H26V15	2	05/07/2017; 05/14/2017

Note: The Landsat ARD tile is not always a complete image tile when only contains one Landsat scene, therefore sometimes the combination of two Landsat scenes is used to get a full image in this tile.

Table 3. The auxiliary variables.

Title	Publication date	Projection system	Accuracy
Annual Maximum Temperature 1981-2010	2012	WGS_1984	0.5 Celsius
Annual Minimum Temperature 1981-2010	2012	WGS_1984	0.5 Celsius
Land Use Land Cover Data Set	2014	GRS 1980	Have not been conducted
USGS National Elevation Data	2019	GRS 1980	1 meter

Ground Data

The ground data is the inventory data for Georgia in 2017 as measured by the Forest Inventory Analysis (FIA) program. The total number of inventory plots was 941 in that year. The criteria we used to filter the volumes in the database are: (1) The inventory year is **2017** (2) The tree status is **Live Tree**. (3) The land basis is **forest land** (4) The estimated element is the **Net volume** (Not blank) (5) The diameter at the breast height is **no less than 0.1 meters**. The final filtered volume is the same characteristic of the net merchantable bole volume of live trees (at least 0.1 meters d.b.h./d.r.c), in cubic meters on forest land. The unit is cubic feet per acre by default, and this is converted to cubic meters per hectare for comparison with other published works.

Volume Estimation

We used *k*NN imputation to estimate the volume using Euclidean distance to find the nearest neighbor. The predicted volume and Euclidean distance are calculated using these formulas:

$$\hat{y}_i = \sum_{j=1,2,3\ldots k}^{k} w_{ij} \times y_{ij} \tag{1}$$

$$d_{ij} = (x_i - x_j)\, W\, (x_i - x_j)' \quad (2)$$

$\hat{y}_i$ is the i_{th} pixel volume in the target set; w_{ij} is the weight assigned to each y_{ij}, equal weight is being used in our study; y_{ij} is the nearest pixel to the i_{th} $(i = 1,2,3, \ldots, n)$ element in the reference set defined by d_{ij}; d_{ij} is the distance used to determine the nearest neighbor; x_i is the i_{th} element in the target set; x_j is the j_{th} $(j = 1,2,3, \ldots, n)$ element in the reference set; W is the inverse of the direct sum of the ancillary variables covariance matrix since we use Euclidean distance (Crookston and Finley, 2008).

The selection of *k* is the *k* value which RMSE is not more than 1% greater than the smallest value of RMSE (McRoberts, 2009).

Accuracy Assessment

The observations were divided into two datasets: the training dataset (708 observations) and the testing dataset (233 observations). The training dataset is used to build the *k*NN learning model, the testing dataset is used to test the model. We used RMSE and Bias to test accuracy.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (3)$$

$$RMSE\ (\%) = \frac{RMSE}{\bar{y}} \times 100 \quad (4)$$

$$Bias = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{n} \quad (5)$$

$$Bias\ (\%) = \frac{Bias}{\bar{y}} \times 100 \quad (6)$$

After testing the model, we will use a best model to create a volume map. Then county-level volume estimates will be extracted from the map. We will compare the standard error of county-level volume estimates with the standard error of county-level FIA estimates at 68% confidence interval standard deviation. The formula used to calculate standard error of volume estimates is as follow:

$$SE = \frac{SD}{\sqrt{n}} \tag{7}$$

The SD is the standard deviation of the volume estimates in a county, *n* is the number of pixels in that county.

Results

Table 4 is the RMSE and Bias of the training model in each group. The *Landsat* group is the control used for comparing against it the results with other groups. The smallest RMSE (%) of *Landsat + All* group indicates the *k*NN training model in this group performs best among our results. The bias shows the percentage of residuals. As the *k*NN imputation is inherently biased because the smallest predicted volume will not smaller than the smallest observed volume and largest predicted volume will not larger than the largest observed volume (McRoberts, 2009), we will accept the bias of the model as long as it is in the acceptable range.

Since a wide range of observed volumes can cause higher RMSE (%) (Kajisa et al., 2008) so this might be one of the factors in our study because the range of observed volumes spans from 0.26 m^3/ha to 606.96 m^3/ha which is larger than the preferable range < 500 m^3/ha mentioned in Kajisa et al. (2008).

The *Landsat+All* comprises 11 variables, while the other groups use 8 variables. The result of using 11 variables has smaller RMSE (%) than using 8 variables in our study is not consistent with the results of a previous study which showed a relatively simple model is preferable (Tanaka et al., 2015). The reason for this inconsistency in results between the current study and a previous study might be due to different ways of using auxiliary variables. Researchers in the previous study did not use the original spectral value of variables directly, as we did in this study. Instead, they derived secondary products from the spatial data. Such processing of the data can remove some noise within the spatial data, which in turn can increase the accuracy of *k*NN imputation. Besides, variables such as maximum temperature, minimum temperature, and land cover map are not published in 2017 (Table 3). Though they are the latest version we can download during the study. Kajisa et al. (2008) used the digital elevation map to calibrate the spectral value in satellite data, while we used elevation as an independent variable.

Table 4. The RMSE and Bias of each group

Group	RMSE		Bias	
	m^3/ha	%	m^3/ha	%
Landsat (7)	110.64	82.18	1.48	1.10
Landsat+All (11)	102.78	77.61	-7.15	-5.40
Landsat+Max_Temp (8)	116.19	85.17	9.06	6.64
Landsat+Min_Temp (8)	119.64	88.26	0.70	0.51
Landsat+Landcover (8)	116.06	85.07	-3.19	-2.34
Landsat+Elevation (8)	118.42	88.19	-0.85	-0.64

Besides the variables employed, there might exist other factors that lead to high relative RMSE. We conducted a linear regression analysis to better understand what causes the large residuals between the observed and predicted volumes. The pattern of the residual distribution (Figure 2) indicates that the overall predicted volume tends to be overestimated. There is some association between the observed volume and residual volume because the correlation is 0.587.

The slope of the solid line indicates there is a positive relationship, and the p-value (< 0.0001) suggests this is a significant regression model. This result is similar to a previous study which also developed a regression model between the observed volume and residuals (Kajisa et al., 2008). The study also reported that when the volume close to the mean, the estimated volume tends to have smaller residuals. Except for the volume, the land area covered by the imputation may also be a significant factor in the results obtained. Some studies mentioned larger areas can have higher accuracy than smaller areas in *k*NN imputation (Fazakas et al., 1999; Tokola et al., 1996; Reese at al., 2002). However, for a state as heterogeneous as Georgia (coastal to mountainous forests), this may not hold.

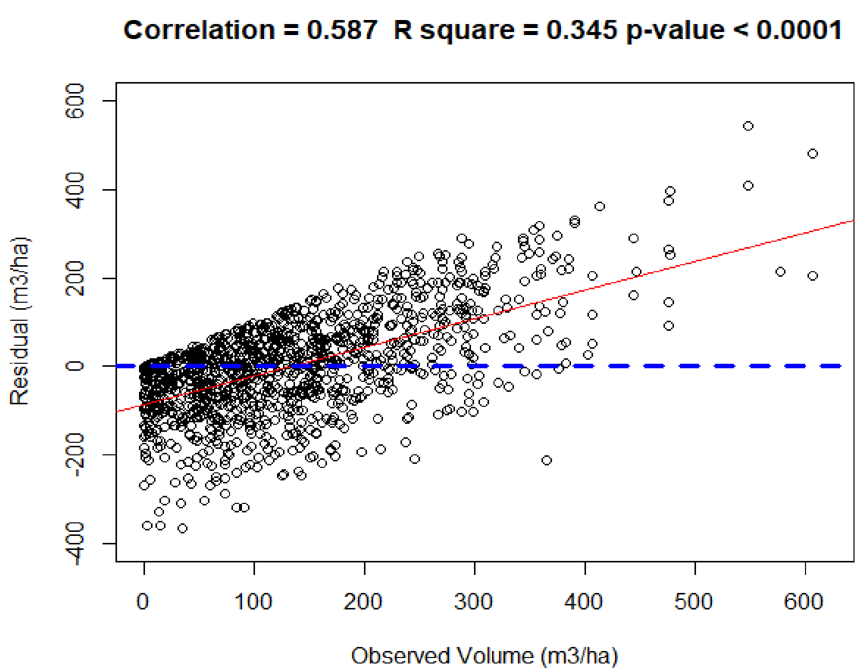

Figure 2. The linear regression of observed volumes versus residuals.
Note: The solid line is the regression line; the dashed line indicates 0 level in residuals.

A volume map was created using Landsat with all auxiliary variables as input with RMSE (%), 77.61% (Figure 3). The high RMSE value in *k*NN imputation is a well-known fact mentioned in many *k*NN studies (Gjertsen, 2007; Tomppo et al., 1999; Fazakas et al., 1999), the range of RMSE (%) among studies ranges from about 44% to 91% (Tanaka et al., 2015; Reese et al., 2002; Mäkelä and Pekkarinen, 2004; Tomppo et al., 2002; Kajisa et al., 2008; Gjertsen, 2007). The map shows the distribution of predicted volume in the state. We can see the majority of the pixels have 100-200 m^3/ha volume (refer as green color), according to the 2017 Georgia forest report published by FIA (Lambert, 2020), we calculated the mean volume per hectare in 2017 to be 130.81 m^3/ha, using the net volume divided by the area of forestland. The volume per hectare from our map is 139.97 m^3/ha. It is reasonable that most pixel volumes on the map belong to 100-200 m^3/ha because 130.81 m^3/ha is inside this range.

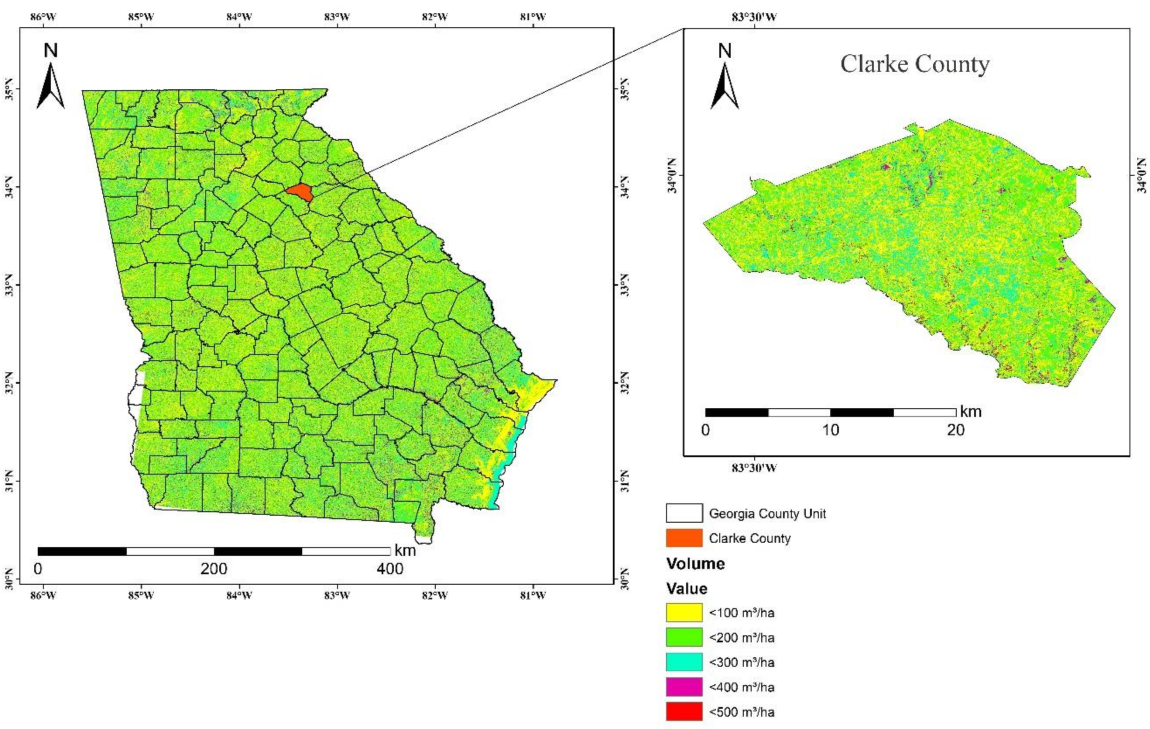

Figure 3. The volume map of Georgia (left) and an example of the volume distribution within a county (right).

Figure 4 is the result of the county-level inventory comparison between the standard error from FIA Evalidator (https://apps.fs.usda.gov/Evalidator/evalidator.jsp) and the standard error of *k*NN. The X-axis is the number of plots used by each county in 2017. The FIA Evalidator shows 2-6 plots were mainly measured for each county in 2017. The small number of plots being used can cause a high standard error, we can see the standard error line (blue) goes down as the number of plot used increases. High standard error indicates the estimation of the county volume from FIA has a large variation. While the standard error line (red) of *k*NN is almost horizontal, and the standard error is not affected by the number of plots measured in each county. Our result is consistent with a previous study which showed the standard error decreased around 10% to 30% percent at the county level (Nilsson et al., 2005). Therefore, we can expect a more reliable county estimation from *k*NN imputation than from FIA.

As the number of pixels is the estimation unit when conducting *k*NN imputation, just like the number of plots used in the FIA inventory, the number of pixels might also affect the standard error. To further understand the relationship between the number of pixels and standard error, we conducted another analysis. Figure 5 suggests there is a strong association ($|\text{correlation}| > 0.5$) between the number of pixels within a county and the standard error of the volume in that county. The *p*-value ($p < 0.0001$) of the regression model indicates the number of pixels has a significant effect on the standard error. The negative relationship between these two variables suggests more pixels within the county, less standard error can be expected. In our result, when the number of pixels exceeds around 1,700,000, the standard error reduction is not as significant as the number of pixels below this point. In general, we can assume the number of pixels in a county is associated with the area of that county. The pixel size in our study is 30 m × 30 m and the area of one pixel is equal to 0.09 hectares. Therefore, when we combine equal or above

1,700,000 pixels' volumes, which is 153,000 hectares, a steady standard error of the estimation can be assumed. Though we are discussing standard error at this point, the results are similar to RMSE as these two both reflect a measure of variation in predicting volume. Reese et al. (2002) also mentioned that using the large area in *k*NN imputation has the potential to provide a more accurate estimation than using the small area. However, here we just simply take the average of large-area volume, and we do not use large-area volumes to conduct *k*NN imputation.

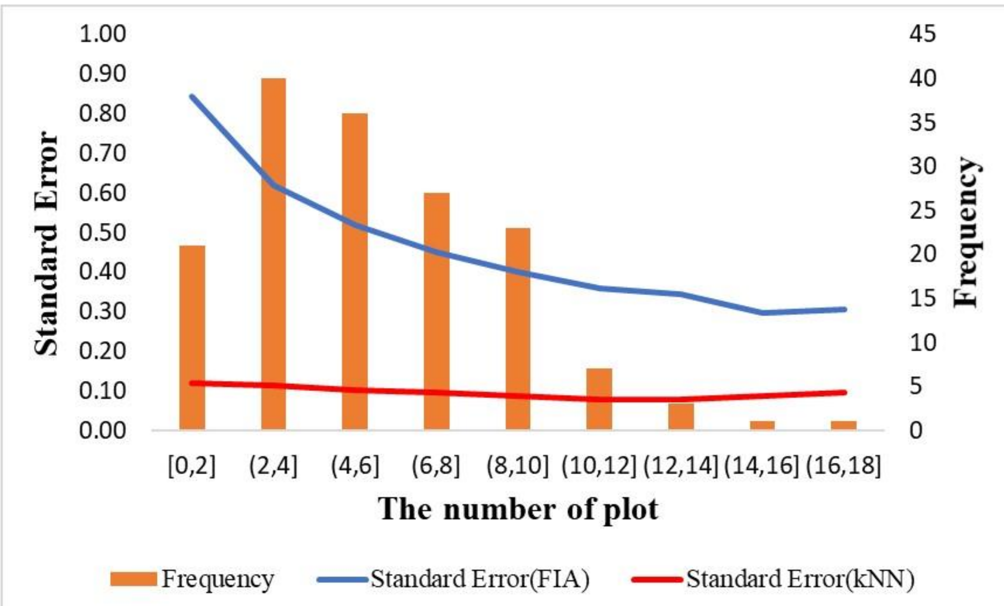

Figure 4. Comparison of sampling error within counties and the frequency of the number of plots used. **Note:** The result is at 68% confidence interval.

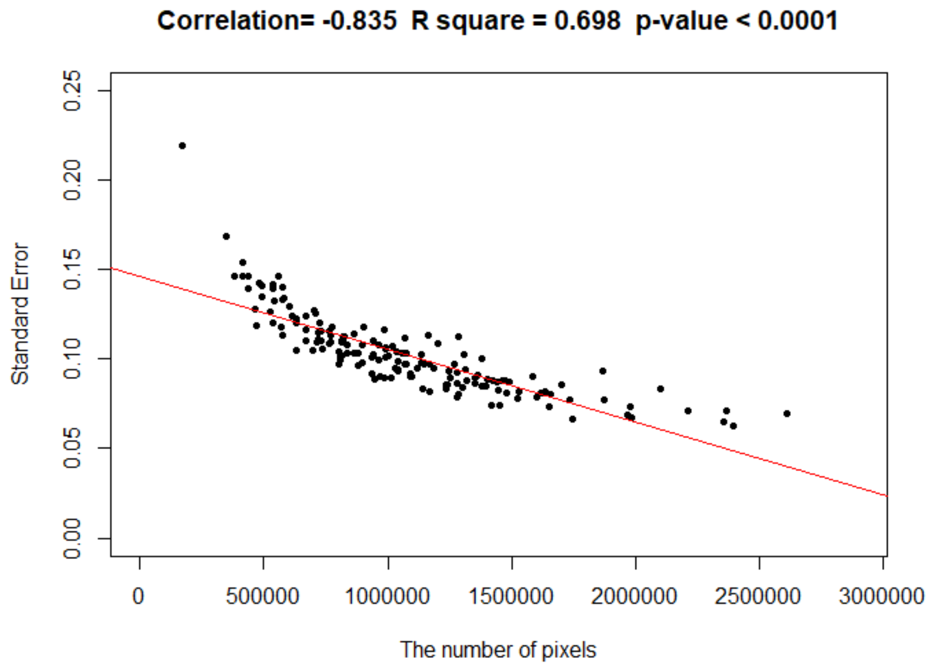

Figure 5. Linear regression of the standard error and the number of pixels within a county. **Note:** The standard error is at 68% confidence interval.

Table 5 is the volume estimates by county. The results in the table are contributions as there is no direct county volume for a single year from the FIA database because only a few plots are measured each year. That is to say, using *k*NN imputation based on spatial data and FIA inventory, we can obtain a time-line and high-resolution inventory. The table shows the smallest total county volume in 2017 is *Clarke* (which is also the smallest county in land area), the highest total county volume is *Ware* (one of the largest counties in the state), the smallest mean volume is *Forsyth*, the highest mean volume is *Echols*; the smallest standard deviation is *Barrow*, the highest standard deviation is *Brantley*. These results are in part directly related to the size of each county.

Table 5. Predicted volume by county.

County code and name	Volume			Area (ha)
	Total (million m³)	Mean (m³/ha)	SD (m³/ha)	
13001 GA Appling	19.89	149.8	107.2	132,752
13003 GA Atkinson	13.45	150.7	105.1	89,249
13005 GA Bacon	10.54	142.3	102.3	74,061
13007 GA Baker	13.80	152.6	102.2	90,410
13009 GA Baldwin	8.45	122.0	96.0	69,258
13011 GA Banks	7.42	122.5	90.7	60,570
13013 GA Barrow	4.90	116.1	**81.3**	42,195
13015 GA Bartow	15.47	127.1	100.6	121,746
13017 GA Ben Hill	8.92	135.7	98.8	65,761
13019 GA Berrien	17.81	150.1	101.3	118,590
13021 GA Bibb	7.57	114.5	94.5	66,129
13023 GA Bleckley	6.91	121.7	95.3	56,752
13025 GA Brantley	17.25	148.8	**127.4**	115,926
13027 GA Brooks	17.45	135.4	105.0	128,914
13029 GA Bryan	16.55	140.4	117.5	117,872
13031 GA Bulloch	25.08	140.5	103.1	178,546
13033 GA Burke	28.97	133.9	97.1	216,435
13035 GA Butts	6.22	127.7	97.6	48,693
13037 GA Calhoun	9.88	134.6	98.6	73,435
13039 GA Camden	29.36	144.8	106.0	202,754
13043 GA Candler	8.54	132.4	92.7	64,503
13045 GA Carroll	17.65	135.3	105.2	130,490
13047 GA Catoosa	5.11	121.5	87.4	42,077
13049 GA Charlton	31.71	156.4	121.4	202,783
13051 GA Chatham	18.86	115.0	87.7	164,064
13053 GA Chattahoochee	8.62	132.5	97.2	65,044
13055 GA Chattooga	10.32	127.1	92.8	81,218
13057 GA Cherokee	15.02	133.5	104.7	112,490
13059 GA Clarke	4.03	128.6	99.4	31,346
13061 GA Clay	6.52	115.8	92.2	56,261
13063 GA Clayton	4.93	132.0	94.1	37,374
13065 GA Clinch	32.11	150.4	109.4	213,473
13067 GA Cobb	10.85	121.6	105.6	89,215
13069 GA Coffee	21.42	137.2	102.2	156,115
13071 GA Colquitt	20.51	142.3	99.4	144,145
13073 GA Columbia	9.47	118.9	90.3	79,715
13075 GA Cook	9.08	150.3	101.6	60,395
13077 GA Coweta	14.16	122.6	89.3	115,452
13079 GA Crawford	9.76	115.4	98.8	84,535

13081 GA Crisp	9.14	125.6	89.7	72,754
13083 GA Dade	5.93	131.5	94.7	45,115
13085 GA Dawson	6.83	123.0	97.2	55,520
13087 GA Decatur	24.12	149.5	103.5	161,374
13089 GA DeKalb	9.79	139.4	103.7	70,211
13091 GA Dodge	16.91	129.8	99.9	130,308
13093 GA Dooly	12.96	126.0	88.9	102,878
13095 GA Dougherty	12.51	144.4	97.4	86,652
13097 GA Douglas	6.65	127.7	106.1	52,059
13099 GA Early	17.25	129.0	100.5	133,722
13101 GA Echols	18.12	**166.3**	119.7	108,984
13103 GA Effingham	19.69	157.2	117.6	125,271
13105 GA Elbert	13.38	138.1	115.5	96,938
13107 GA Emanuel	22.29	124.6	94.4	178,855
13109 GA Evans	6.69	138.0	103.7	48,445
13111 GA Fannin	14.58	143.7	100.1	101,498
13113 GA Fayette	6.82	132.0	89.5	51,615
13115 GA Floyd	17.35	129.2	106.9	134,286
13117 GA Forsyth	6.95	**108.6**	105.7	64,003
13119 GA Franklin	9.31	135.0	101.3	69,009
13121 GA Fulton	17.79	128.5	101.4	138,365
13123 GA Gilmer	14.16	126.8	95.6	111,672
13125 GA Glascock	4.68	125.0	99.4	37,425
13127 GA Glynn	21.98	144.8	104.3	151,791
13129 GA Gordon	11.02	118.8	96.7	92,724
13131 GA Grady	17.64	148.0	108.0	119,184
13133 GA Greene	13.14	124.9	105.3	105,209
13135 GA Gwinnett	14.23	125.8	100.1	113,102
13137 GA Habersham	8.87	122.7	86.8	72,259
13139 GA Hall	12.10	108.8	95.3	111,167
13141 GA Hancock	15.49	124.9	100.0	123,989
13143 GA Haralson	11.10	151.4	90.8	73,344
13145 GA Harris	15.81	129.0	104.2	122,485
13147 GA Hart	8.04	120.8	94.1	66,556
13149 GA Heard	9.42	120.8	95.6	77,983
13151 GA Henry	10.74	127.0	88.7	84,558
13153 GA Houston	12.24	124.4	94.5	98,384
13155 GA Irwin	13.10	139.5	101.1	93,953
13157 GA Jackson	10.93	123.0	89.0	88,851
13159 GA Jasper	12.52	129.5	107.2	96,712
13161 GA Jeff Davis	12.11	139.4	105.8	86,895
13163 GA Jefferson	17.02	124.0	96.5	137,246
13165 GA Jenkins	11.77	128.8	90.5	91,368
13167 GA Johnson	9.41	118.5	96.9	79,431

13169 GA Jones	13.05	127.4	104.8	102,400
13171 GA Lamar	6.08	126.3	87.8	48,121
13173 GA Lanier	7.87	152.1	101.4	51,749
13175 GA Laurens	26.88	126.8	99.2	212,012
13177 GA Lee	12.94	138.1	95.5	93,669
13179 GA Liberty	22.43	143.5	111.9	156,277
13181 GA Lincoln	7.59	113.9	90.3	66,668
13183 GA Long	16.37	156.3	122.0	104,702
13185 GA Lowndes	19.30	145.9	106.5	132,236
13187 GA Lumpkin	9.08	123.4	91.8	73,592
13189 GA McDuffie	12.35	117.4	88.7	105,131
13191 GA McIntosh	9.70	131.1	99.0	73,966
13193 GA Macon	12.09	127.1	106.3	95,160
13195 GA Madison	8.12	117.7	95.1	69,022
13197 GA Marion	20.64	138.7	102.7	148,866
13199 GA Meriwether	15.93	121.7	89.4	130,881
13201 GA Miller	11.03	150.2	99.9	73,459
13205 GA Mitchell	19.61	147.4	98.8	133,053
13207 GA Monroe	12.82	124.4	103.8	103,013
13209 GA Montgomery	8.75	138.1	106.7	63,399
13211 GA Morgan	12.26	133.5	108.1	91,837
13213 GA Murray	11.13	123.9	100.9	89,784
13215 GA Muscogee	7.11	124.2	97.1	57,235
13217 GA Newton	9.09	125.7	93.5	72,299
13219 GA Oconee	6.35	131.6	102.1	48,266
13221 GA Oglethorpe	15.27	133.4	109.4	114,493
13223 GA Paulding	10.74	132.0	112.1	81,406
13225 GA Peach	4.45	113.5	96.4	39,180
13227 GA Pickens	7.69	127.6	95.3	60,261
13229 GA Pierce	13.38	150.5	115.6	88,954
13231 GA Pike	7.08	124.7	83.5	56,815
13233 GA Polk	12.32	152.4	102.3	80,846
13235 GA Pulaski	8.70	133.7	94.5	65,056
13237 GA Putnam	11.55	123.6	106.4	93,411
13239 GA Quitman	5.25	126.2	90.1	41,584
13241 GA Rabun	13.50	138.3	100.7	97,605
13243 GA Randolph	13.14	117.7	92.3	111,582
13245 GA Richmond	11.25	132.1	97.9	85,167
13247 GA Rockdale	4.15	121.4	90.0	34,206
13249 GA Schley	5.27	121.4	98.9	43,458
13251 GA Screven	25.33	149.0	106.3	170,075
13253 GA Seminole	8.73	131.5	102.6	66,442
13255 GA Spalding	6.55	126.7	85.8	51,683
13257 GA Stephens	5.59	117.1	91.6	47,705

13259 GA Stewart	16.01	133.4	103.1	120,063
13261 GA Sumter	14.62	114.6	88.2	127,571
13263 GA Talbot	13.71	134.1	109.2	102,225
13265 GA Taliaferro	6.78	134.0	109.9	50,602
13267 GA Tattnall	18.11	143.1	105.0	126,580
13269 GA Taylor	11.53	117.3	96.1	98,322
13271 GA Telfair	15.89	138.2	105.0	115,013
13273 GA Terrell	10.58	121.0	89.1	87,449
13275 GA Thomas	22.46	157.1	113.8	142,997
13277 GA Tift	9.46	135.9	100.0	69,608
13279 GA Toombs	12.95	134.7	100.7	96,138
13281 GA Towns	6.01	135.0	99.0	44,536
13283 GA Treutlen	6.93	132.2	102.3	52,453
13285 GA Troup	13.55	117.3	97.9	115,499
13287 GA Turner	9.47	126.2	94.2	75,094
13289 GA Twiggs	11.94	127.1	96.5	93,914
13291 GA Union	12.53	147.1	107.3	85,195
13293 GA Upson	10.71	126.3	99.7	84,819
13295 GA Walker	16.35	141.2	91.3	115,787
13297 GA Walton	10.40	121.7	86.3	85,456
13299 GA Ware	**38.40**	163.2	112.0	235,281
13301 GA Warren	9.57	128.8	102.0	74,272
13303 GA Washington	22.35	126.1	96.5	177,284
13305 GA Wayne	26.24	156.0	127.2	168,193
13307 GA Webster	6.81	125.0	100.5	54,449
13309 GA Wheeler	10.66	137.0	106.4	77,789
13311 GA White	7.94	126.7	87.3	62,721
13313 GA Whitfield	10.43	138.4	98.7	75,378
13315 GA Wilcox	12.81	129.5	94.9	98,962
13317 GA Wilkes	16.72	136.1	106.1	122,818
13319 GA Wilkinson	15.36	131.2	95.6	117,056
13321 GA Worth	19.39	130.3	94.7	148,834

Note: The number with red color indicates the maximum value in that column; the blue color indicates the minimum value in that column; the SD is the standard deviation.

Discussion

The smallest RMSE (%) of *k*NN model in our study is 77.61% which is higher than the previous studies (Gjertsen 2007; Gu et al., 2006; Nilsson et al., 2005). The reasons might be: First, we did not register the Landsat image with digital elevation data (Kajisa et al., 2008). Especially Georgia is a heterogeneous state (coastal to mountainous forests), therefore, this reason might dramatically affect the accuracy. Second, minimum and maximum temperature maps might not suitable to be used directly, some calculations or extractions before using the data can reduce RMSE (%) (Tanaka et al., 2015). Finally, the range of the observed volume in the model is also a potential factor. As the range in our data is large (0.26 m^3/ha to 606.96 m^3/ha), we can explore whether only use the observed volumes which are close to mean will have lower RMSE (%) in the future.

Though the RMSE (%) in our study is high, we display *k*NN imputation has the potential to predict more accurate results if some improvements can be made. The most attractive advantage of *k*NN imputation is to obtain a high-resolution, common-timeline inventory that can be used in forest management. Future research can test whether using multiple spectral data such as a combination of satellite imagery and lidar can achieve better results in kNN imputation.

Conclusion

In attempting to determine an adequate map of timber volume across a broad area such as the State of Georgia, adding auxiliary variables has the potential of reducing RMSE (%) in a *k*NN model, but how to add these variables to obtain better results needs to be considered. Overall, the best *k*NN model in our study is using Landsat data with maximum temperature map, minimum temperature map, land cover map, and elevation map with a 77.61% RMSE and -5.4% bias.

REFERENCES

Bechtold, W.A., & Patterson, P. L. (2005). *The enhanced forest inventory and analysis program - national sampling design and estimation procedures* (SRS-GTR-80). U.S. Department of Agriculture, Forest Service, Southern Research Station, Asheville, NC.

Crookston, N. L., & Finley, A. O. (2008). yaImpute: an R package for kNN imputation. *Journal of Statistical Software. 23 (10).* 16 p.

Dwyer, J. L., Roy, D. P., Sauer, B., Jenkerson, C. B., Zhang, H. K., & Lymburner, L. (2018). Analysis ready data: enabling analysis of the Landsat archive. *Remote Sensing*, *10*(9), 1363.

Eskelson, B. N., Temesgen, H., Lemay, V., Barrett, T. M., Crookston, N. L., & Hudak, A. T. (2009). The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. *Scandinavian Journal of Forest Research*, *24*(3), 235-246.

Fazakas, Z., Nilsson, M., & Olsson, H. (1999). Regional forest biomass and wood volume estimation using satellite data and ancillary data. *Agricultural and Forest Meteorology*, *98*, 417-425.

Franco-Lopez, H., Ek, A. R., & Bauer, M. E. (2001). Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. *Remote Sensing of Environment*, *77*(3), 251-274.

Ge, G., Shi, Z., Zhu, Y., Yang, X., & Hao, Y. (2020). Land use/cover classification in an arid desert-oasis mosaic landscape of China using remote sensed imagery: Performance assessment of four machine learning algorithms. *Global Ecology and Conservation*, *22*, e00971.

Gjertsen, A. K. (2007). Accuracy of forest mapping based on Landsat TM data and a kNN-based method. *Remote Sensing of Environment*, *110*(4), 420-430.

Gu, H., Dai, L., Wu, G., Xu, D., Wang, S., & Wang, H. (2006). Estimation of forest volumes by integrating Landsat TM imagery and forest inventory data. *Science in China Series E: Technological Sciences*, *49*(1), 54-62.

Kajisa, T., Murakami, T., Mizoue, N., Kitahara, F., & Yoshida, S. (2008). Estimation of stand volumes using the k-nearest neighbors method in Kyushu, Japan. *Journal of Forest Research*, *13*(4), 249-254.

Lambert, S. (2020). *Georgia's forests, 2017: Annual update. Resource Update* (FS–RU-226). U.S. Department of Agriculture Forest Service, Southern Research Station, Asheville, NC.

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*(2), 109-119.

Mäkelä, H., & Pekkarinen, A. (2004). Estimation of forest stand volumes by Landsat TM imagery and stand-level field-inventory data. *Forest Ecology and Management*, *196*(2-3), 245-255.

McRoberts, R. E. (2009). Diagnostic tools for nearest neighbors techniques when used with satellite imagery. *Remote Sensing of Environment*, *113*(3), 489-499.

McRoberts, R. E. (2012). Estimating forest attribute parameters for small areas using nearest neighbors techniques. *Forest Ecology and Management*, *272*, 3-12.

McRoberts, R. E., Nelson, M. D., & Wendt, D. G. (2002). Stratified estimation of forest area using satellite imagery, inventory data, and the k-Nearest Neighbors technique. *Remote Sensing of Environment*, *82*(2-3), 457-468.

Meng, Q., Cieszewski, C. J., Madden, M., & Borders, B. E. (2007). K nearest neighbor method for forest inventory using remote sensing data. *GIScience & Remote Sensing*, *44*(2), 149-165.

Moeur, M., & Stage, A. R. (1995). Most similar neighbor: an improved sampling inference procedure for natural resource planning. *Forest Science*, *41*(2), 337-359.

National Aeronautics and Space Administration. (2020). *A Landsat Timeline*. National Aeronautics and Space Administration, Goddard Space Flight Center, Greenbelt, MD. https://landsat.gsfc.nasa.gov/a-landsat-timeline/

Nguyen, T. H., Jones, S., Soto-Berelov, M., Haywood, A., & Hislop, S. (2018). A comparison of imputation approaches for estimating forest biomass using landsat time-series and inventory data. *Remote Sensing*, *10*(11), 1825.

Nilsson, M., Holm, S., Reese, H., Wallerman, J., & Engberg, J. (2005). Improved forest statistics from the Swedish national forest inventory by combining field data and optical satellite data using post-stratification. *ForestSat 2005*, 22.

Ohmann, J. L., & Gregory, M. J. (2002). Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, USA. *Canadian Journal of Forest Research*, *32*(4), 725-741.

Reese, H., Nilsson, M., Sandström, P., & Olsson, H. (2002). Applications using estimates of forest parameters derived from satellite and forest inventory data. *Computers and Electronics in Agriculture*, *37*(1-3), 37-55.

Stage, A. R., & Crookston, N. L. (2007). Partitioning error components for accuracy-assessment of near-neighbor methods of imputation. *Forest Science*, *53*(1), 62-72.

Tanaka, S., Takahashi, T., Nishizono, T., Kitahara, F., Saito, H., Iehara, T., ... & Awaya, Y. (2015). Stand volume estimation using the k-NN technique combined with forest inventory data, satellite image data and additional feature variables. *Remote Sensing*, *7*(1), 378-394.

Tokola, T., Pitkänen, J., Partinen, S., & Muinonen, E. (1996). Point accuracy of a non-parametric method in estimation of forest characteristics with different satellite materials. *International Journal of Remote Sensing*, *17*(12), 2333-2351.

Tompalski, P., White, J. C., Coops, N. C., & Wulder, M. A. (2019). Demonstrating the transferability of forest inventory attribute models derived using airborne laser scanning data. *Remote Sensing of Environment*, *227*, 110-124

Tomppo, E. R. K. K. I., & Katila, M. (1991). Satellite image-based national forest inventory of finland for publication in the igarss' 91 digest. In *[Proceedings] IGARSS'91 Remote Sensing: Global Monitoring for Earth Management*: *Vol. 3* (pp. 1141-1144). IEEE.

Tomppo, E., Nilsson, M., Rosengren, M., Aalto, P., & Kennedy, P. (2002). Simultaneous use of Landsat-TM and IRS-1C WiFS data in estimating large area tree stem volume and aboveground biomass. *Remote Sensing of Environment*, *82*(1), 156-171.

Tomppo, E., Olsson, H., Ståhl, G., Nilsson, M., Hagner, O., & Katila, M. (2008). Combining national forest inventory field plots and remote sensing data for forest databases. *Remote Sensing of Environment*, *112*(5), 1982-1999.

Trotter, C. M., Dymond, J. R., & Goulding, C. J. (1997). Estimation of timber volume in a coniferous plantation forest using Landsat TM. *International Journal of Remote Sensing*, *18*(10), 2209-2223.

United States Department of Agriculture, Natural Resources Conservation Service. (2019). Geospatial Data Gateway. United States Department of Agriculture, Natural Resources Conservation Service, Washington, DC. https://datagateway.nrcs.usda.gov/GDGOrder.aspx?order=QuickState

United States Forest Service. (2012, March 15). *Historian Lincoln Bramwell on History of the Forest Service* [Video]. YouTube. https://www.youtube.com/watch?v=EPRDHe9ZTaA&feature=emb_logo

Yazdani, M., Jouibary, S. S., Mohammadi, J., & Maghsoudi, Y. (2020). Comparison of different machine learning and regression methods for estimation and mapping of forest stand attributes using ALOS/PALSAR data in complex Hyrcanian forests. *Journal of Applied Remote Sensing*, *14*(2), 024509.