IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 14, 2019, accepted September 13, 2019, date of publication September 17, 2019, date of current version October 1, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2941966

Sparkpr: An Efficient Parallel Inversion of Forest Canopy Closure

GUANGSHENG CHEN[1], TONGTONG LOU[1], WEIPENG JING[1], AND ZEYU WANG[2]
[1]College of Information and Computer Engineering, Northeast Forestry University, Harbin 150040, China
[2]School of Information and Intelligent Engineering, Sanya University, Sanya 572022, China

Corresponding author: Weipeng Jing (weipeng.jing@outlook.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 31770768, in part by the Natural Science Foundation of Heilongjiang Province of China under Grant F2017001, in part by the Heilongjiang Province Applied Technology Research and Development Program Major Project under Grant GA18B301, and in part by the China State Forestry Administration Forestry Industry Public Welfare Project under Grant 201504307.

ABSTRACT Forest canopy closure is an important parameter to study forest ecosystem and understand the status of forest resources. With the development of remote sensing big data, the amount of remote sensing data has increased sharply, which makes the existing serial processing of remote sensing data face severe challenges.In order to satisfy the requirements of efficient remote sensing data processing, Spark open source framework is applied to the parallel processing of remote sensing images, and a parallel forest canopy density inversion algorithm based on Spark is proposed. We call this algorithm Sparkpr. Based on the GF-1 remote sensing images and 80 actual measured sample points obtained by Maoershan Laoshan Experimental Forest Farm of Northeast Forestry University in 2016. In this paper, a multi-element linear regression algorithm is used to carry out parallel inversion of the forest canopy density in the Laoshan Experimental Forest Farm of the Maoershan. The comparison experiment between single machine mode and spark standalone and spark on yarn mode is carried out. The experimental results show that the serial and parallel inversion results of forest depression density based on the model are consistent, and the parallel inversion results are accurate and credible. With the increase of computing nodes, the efficiency of parallel inversion is also improving.

INDEX TERMS Cloud computing, spark model, parallel computing, forest crown closure.

I. INTRODUCTION

Forest is the largest ecosystem on land. It covers a large area, distributes widely, is rich in materials and has complex structure. It not only provides valuable resources for human survival and development, but also plays an important role in protecting the earth's ecosystem and water resources [1]. Forest Crown Closure is the radio of the canopy projection area to the woodland area. It is one of the important characteristic factors of forest ecosystem state [2] and environmental evaluation index and now it is widely used in evaluation of forestry, judgment of urban climate and heat island effect [3]. Forest crown closure, as an important index in forest survey, is an important index to reflect the spatial structure of ecosystem and tree stand density [4]. It plays an important role in forest resources inventory.

It is of great significance to study how to efficiently extract the forest. Vegetation index NDVI, RVI, DVI, ARVI and spectral combination extracted from 80 measured data samples and GF-1 remote sensing images of Laoshan experimental forest farm in Maoershan, Northeast Forestry University in 2016 are used as experimental sources in this paper. A parallel inversion method of forest canopy density based on spark on yarn and Python Numpy was proposed by using multivariate linear regression method of remote sensing inversion.

The associate editor coordinating the review of this manuscript and approving it for publication was El-Hadi M. Aggoune.

II. RELATED WORK

Traditional forest crown closure measurement method has visual method, crown projection method, spline method, sample point method and statistical method [5]. Li *et al.* [6] measured the value of depression density by crown projection. However, these traditional measurement methods are not accurate enough. They take a lot of manpower, material and time to estimate forest crown closure on a large scale [7]. Moreover, the data obtained by these methods are only based on point-scale data, which is not conducive to the measurement of large scale and large scale space [8]. With the development of science and technology, a method of forest

crown closure measurement based on remote sensing data comes out [9].

Among them, the development of traditional lidar technology and optical remote sensing is earlier. Iqbal *et al.* [10] used Glas lidar data to estimate forest crown closure, Moeser *et al.* [11] synthesized hemispheric photos by vehicle lidar, and estimated the canopy density and leaf area index of the forest. However, due to the discontinuous distribution of GLAS laser spots in space, seamless coverage can not be achieved. The development of remote sensing and high spectrum technology made up for this deficiency. At present, many studies are based on the regression of the canopy closure values of Landsat, LandsatTM and other images in a certain band or combination of different bands with the observed canopy closure values of sample points, and establish models to predict canopy closure values in a wide range of areas [12]. Xuecheng and Cai [13] established the relationship model between each gray band of remote sensing image and forest canopy density by using CMB resource satellite. Zhang *et al.* [14] combined with airborne LiDAR and LANDSAT ETM data to extract vegetation index, and multiple stepwise regression, random forest and Cubist models were used to estimate forest crown closure. The highest precision of inversion was up to 79.883%. Niu *et al.* [15] based on the TM image data and linear regression method to forecast the forest crown closure of Ban Chengzi. But the inversion model established is not accurate enough because only one variable of NDVI is used for regression analysis. Pu *et al.* [16] based on Landsat TM data to forecast the forest crown closure of oak with unconstrained least squares method and artificial neural network model. But the inversion accuracy was not very high due to the low resolution of Landsat TM. And then Pu *et al.* [17] used stepwise regression method based on Hyperion image to select variables closely related to forest crown closure to model multivariate regression in order to estimated the crown closure of the forest. The inversion accuracy was 85%. The GF-1 satellite launched by China took into account the characteristics of high resolution, high space, high width and high time which compared with poor penetration of Landsat series data, easy saturation [18], low resolution, low band of high spectrum data and the disadvantages of mixed pixel problem [19]. In principle, the GF-1 satellite can break through the dependence of domestic researchers on Landsat satellite and high spectrum data [20].

However, with the development of remote sensing technology, the traditional methods of collecting and analyzing remote sensing images are also changing. Large-scale remote sensing data are produced every day [21], and the amount of remote sensing data is increasing, which makes the existing serial remote sensing data processing face severe challenges. The efficiency of data extraction and calculation of forest crown closure needs to be improved urgently. In order to satisfy the high efficiency and high speed processing of remote sensing data, researchers at home and abroad combine parallel computing into the calculation of remote sensing data. Meng [22] put forward a method of research on normalized vegetation index of remote sensing image based on GPU, which proved that there is a good speedup ratio in information extraction of normalized vegetation index. But the expansion of the model was not good enough to deal with a large number of remote sensing information processing. Fu *et al.* [23] applied Map Reduce framework to biomass inversion, and improved the efficiency of inversion. Li *et al.* [24] made Map Reduce applied to the processing of remote sensing data, which proved the improvement of data processing efficiency. However, MapReduce frequently accessed I/O to read data, which limited the speed of parallel processing. Compared with MapReduce, Spark provided a better mechanism for data reuse, strong data set abstraction, and made up for the deficiency of MapReduce. Greatly improve the data processing efficiency Greatly improve the data processing efficiency [25].

In summary, a parallel inversion algorithm of forest depression density based on Spark is proposed in this paper. We call this algorithm Sparkpr. The serial inversion is compared with the parallel inversion under spark-standalone model and spark on yarn model, and the accuracy, efficiency and performance of depression density inversion under this method are evaluated.

III. DATA STORAGE BASED ON HDFS

Spark is a cloud computing platform designed by AMPLab laboratory at the university of California, Berkeley in 2009, which summarized the insufficient development of Hadoop. The Spark is a computing framework based on memory. Its core data structure is Resilient Distributed Dataset. We can regard RDD in Spark as a transformation method under certain conditions and circumstances. Spark can convert the data set of original remote sensing images into RDD file format and save the processed results in memory. The corresponding program is then executed, the relative results are calculated, and the results are written in HDFS.

Distributed File System HDFS (Hadoop Distribute File System) is used to store the underlying data in Hadoop system and access the data sets in distributed file system by streaming [26]. The data storage designed and implemented is based on the distributed file system (HDFS), then read remote sensing data to Spark cluster for data processing. Since the data after parallel inversion is still stored in HDFS, the storage structure of HDFS is divided into two aspects: original data storage and parallel inversion data storage with canopy closure.

The advantage of distributed storage system is that it can build a high efficiency and high throughput parallel framework on a general system framework. In order to improve the speed of data processing, the remote sensing image is cut into several blocks and stored in a distributed file system in the form of each small file. We adopt flat belt block mode on the remote sensing image of remote sensing image after image mosaic and the RBC of block processing, because we only need to care about the pixels of remote sensing images in the density inversion [27]. The remote sensing data is stored

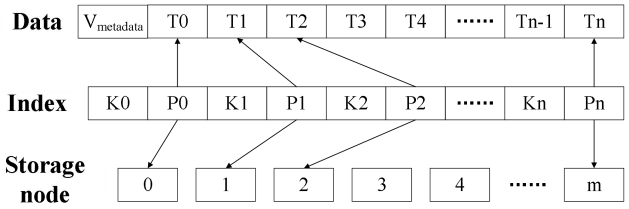

FIGURE 1. Data storage model graph of image block.

on HDFS by means of local data storage [28]. The remote sensing data stored on HDFS is read by GDAL, and the remote sensing image is divided into blocks. The default size of the blocks is 64MB.

When storing data, a request is first made to the name-node master node to transmit remote sensing data to the node. Combined with the characteristics of distributed data storage, the storage structure of key value is used to store the index queue through Map-file, and Index is used as the data index of files, which is mainly used to record the key value of each Record. As shown in figure 1. Combined key-value pair, K_x is image block coding and T_x is image block binary stream. The index queue and data queue are used to store K_x and T_x, which can realize fast retrieval based on K value. V_{metadata} in the data queue represent meta data in the data, and the value of P are position pointers pointing to V_{metadata} of data. Map-file serialization stores the encoding of image blocks in key, the image information in value, and the information after partitioning on each information node to realize the storage of data and index. The image blocks in the data queue will be stored in the storage nodes 0 to m, and the data blocks will be stored on each node according to the multiple replicas, which ensures the high concurrency and availability of the system.

The remote sensing image is first read from the HDFS in the space of the request in parallel inversion. Then link to the HDFS cluster after reading sensing images. O pen the index file in Map-file, The index file in Map-file is loaded into memory, and quickly navigate from memory to where the file is, read the binary stream of the image block, parse the binary stream, and locate the coordinates of the data block. After transforming the blocks into < key, value > pairs, the blocks are distributed to each node for parallel algorithm calculation in Spark, and the results are stored in HDFS.

IV. INVERSION MODEL BASED ON MULTIVARIATE LINEAR REGRESSION

Multivariate linear regression, which is an important method in traditional analysis and data mining, is to predict or estimate dependent variables by the optimal combination of multiple independent variables [29]. It is widely used in all fields of science. It is possible to test and analyze the significance of the influence of each independent variable on the dependent variable [30]. Multiple linear regression algorithm is highly accurate in forest canopy density inversion and is recognized by many scholars at home and abroad.. Only the more significant independent variables are selected to establish the optimal multiple linear regression equation. In this paper, the establishment of inversion model is divided into two parts: the selection of relevant factors and the establishment of model.

A. SELECTION OF CORRELATION FACTORS

Firstly, 32 spectral combinations of vegetation index NDVI, RVI, IPVI, GNDVI, GBNDVI, EVI, DVI, ARVI and high-resolution bands of ρG, ρB, ρR, ρNIR are extracted from GF-1 remote sensing images.

Its index is described as follows:

(1)Normalized vegetation index (NDVI): One of the important indicators reflecting forest growth and nutrition information can be used to detect vegetation growth status vegetation coverage and eliminate some radiation errors which can reflect the background effect of plant canopy. The expression is:

$$NDVI = \frac{\rho NIR - \rho R}{\rho NIR + \rho R} \tag{1}$$

(2)Ratio vegetation Index (RVI): It is a sensitive indicator of green plants. The expression is:

$$RVI = \frac{\rho NIR}{\rho R} \tag{2}$$

(3)Near-infrared percentage vegetation index (IPVI): Reflects the density of vegetation. The expression is:

$$IPVI = \frac{\rho NIR}{\rho NIR + \rho R} \tag{3}$$

(4)The green normalized vegetation index (GNDVI): Refers to the normalized vegetation index in the green band. The expression is:

$$GNDVI = \frac{\rho NIR - \rho G}{\rho NIR + \rho G} \tag{4}$$

(5)Normalized green difference vegetation index (GBNDVI): Improved normalized vegetation index in blue and green bands based on GNDVI. The expression is:

$$GBNDVI = \frac{[\rho NIR - (\rho B + \rho G)]}{\rho NIR + \rho R + \rho G} \tag{5}$$

(6)Enhanced vegetation index (EVI): by adding a blue band to enhance the vegetation signal, correct the effects of soil background and aerosol scattering. The expression is:

$$EVI = 2.5 \times \frac{\rho NIR - \rho R}{\rho NIR + 6 \times \rho R - 7.5 \times \rho B + 1} \tag{6}$$

(7)The difference vegetation index (DVI): It can well reflect the change of vegetation coverage, especially sensitive to soil background change. The expression is:

$$DVI = \rho NIR - \rho R \tag{7}$$

(8)The atmospheric-resistance vegetation index (ARVI): Kaufman [31] developed ARVI in 1992 to reduce the changes in the atmospheric vegetation index due to atmospheric reasons. The expression is:

$$DVI = \rho NIR - \rho R \tag{8}$$

(1) The 24 spectral combination values: $\rho B+\rho G$, $\rho B+\rho R$, $\rho B+\rho NIR$, $\rho G+\rho NIR$, $\rho G+\rho R$, $\rho NIR+\rho R$, $\rho B\times\rho G$, $\rho B\times\rho R$, $\rho B\times\rho NIR$, $\rho G\times\rho NIR$, $\rho G\times\rho R$, $\rho NIR\times\rho R$, $\frac{\rho B}{SUM}$, $\frac{\rho G}{SUM}$, $\frac{\rho R}{SUM}$, $\frac{\rho NIR}{SUM}$, $\frac{\rho R\times\rho NIR}{\rho G}$, $\frac{\rho G+\rho NIR+\rho R}{\rho B}$, $\frac{\rho R\times\rho G}{\rho B}$, $\frac{\rho G\times\rho B}{\rho N}$, $\frac{\rho B\times\rho N}{\rho R}$, $\frac{\rho G+\rho B+\rho R}{\rho NIR}$, $\frac{\rho B+\rho R+\rho NIR}{\rho G}$, $\frac{\rho G+\rho NIR+\rho B}{\rho R}$。

In the above formula, ρB is the reflectivity of blue light band, ρG is the reflectivity of green light band, ρR is the reflectivity of red light band, ρNIR is the reflectivity of near infrared band, SUM is the sum of the above four-band values. From the above 32 variables, five parameters with the highest correlation coefficient are extracted as independent variables to establish the inversion mode of forest canopy densityl.

The formula of correlation coefficient(η) is:

$$\eta = \frac{\mathrm{Cov}\,(X, Y)}{\sqrt{V\mathrm{ar}(X)\cdot Var(Y)}} \tag{9}$$

Of which, $C\mathrm{ov}(X, Y)$ is co-variance. The expression is:

$$C\mathrm{ov}(X, Y) = E((X-\mu)(Y-\nu)) = E(X, Y) - \mu\nu \tag{10}$$

$V\mathrm{ar}(X)$ is Data dispersion. The expression is:

$$V\mathrm{ar}(X) = E\left[(X-\mu)^2\right] = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}(X_i-\mu)^2 \tag{11}$$

In the above formula, μ means $E\,(X)$, expectation of X. ν means $E\,(Y)$, expectation of Y.

B. ESTABLISHMENT OF MODEL BASED ON MULTIVARIATE LINEAR REGRESSION

In this paper, multiple linear regression algorithm is used to establish the inversion model of forest depression density. The multivariate linear regression model mainly examines the correlation between variables, assuming that variable Y and variables X_1, X_2, X_3 have a linear relationship, then their linear regression model can be expressed as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \cdots + \beta_n X_n + \mu \tag{12}$$

Y is a dependent variable, and x is n general variables that can be accurately measured and controllable are called independent variables. For observation of the number i:

$$\begin{aligned} Y_i = {} & \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} \cdots + \beta_n X_{ni} \\ & + \mu i, \quad i = 1, 2, 3\cdot\cdot, n \\ & \times \begin{cases} Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \beta_3 X_{31} \\ \quad + \cdots + \beta_{\mathrm{n}} X_{n1} + \mu 1 \\ Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \beta_3 X_{32} \\ \quad + \cdots + \beta_{\mathrm{n}} X_{n2} + \mu 2 \\ Y_3 = \beta_0 + \beta_1 X_{13} + \beta_2 X_{23} + \beta_3 X_{33} \\ \quad + \cdots + \beta_{\mathrm{n}} X_{n3} + \mu 3 \\ \vdots \\ Y_k = \beta_0 + \beta_1 X_{1k} + \beta_2 X_{2k} + \beta_3 X_{3k} \\ \quad + \cdots + \beta_{\mathrm{n}} X_{nk} + \mu k \end{cases} \end{aligned} \tag{13}$$

Estimation of parameter values of $\beta_1, \beta_2, \beta_3 \cdots \beta_n$ by least square method. Make the sum of squares of deviations (S) between the i-th observation value Y_{i} and the corresponding function value $\hat{Y}_{\mathrm{i}}$ is minimized.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \cdots + \beta_n X_n + \mu \tag{14}$$

It can be concluded by sorting it out:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum X_{1i} + \hat{\beta}_2 \sum X_{2i} + \cdots \\ \quad + \hat{\beta}_k \sum X_{ki} = \sum Y_i \\ \hat{\beta}_0 + \hat{\beta}_1 \sum {X^2}_{1i} + \hat{\beta}_2 \sum {X^2}_{2i} + \cdots \\ \quad + \hat{\beta}_k \sum X_{ki} X_{1i} = \sum X_{1i} Y_i \\ \vdots \\ \hat{\beta}_0 + \hat{\beta}_1 \sum X_{1i} X_{\mathrm{ki}} + \hat{\beta}_2 \sum X_{2i} X_{ki} + \\ \cdots \hat{\beta}_k \sum X_{ki} X_{ki} = \sum X_{ki} Y_i \end{cases} \tag{15}$$

The value of $\beta_0, \beta_1, \beta_2, \beta_3, \cdots, \beta_{\mathrm{n}}$ can be obtained by solving the matrix equation.

Because only five independent variables and one constant value were selected in the model. So we should only find out the value of β_0, β_1, β_2, β_3, β_4, β_5. The multi-source linear regression inversion model of forest canopy density can be obtained.

V. PARALLEL INVERSION OF FOREST CROWN CLOSURE BASED ON SPARK

In this paper, the parallel inversion of forest depression density in Laoshan Experimental Forest Farm of Maoershan Mountain based on multiple linear regression algorithm is studied to test the processing speed of remote sensing big data based on Spark. On the basis of multivariate linear regression algorithm, the independent variables with high linear influence of dependent variables are analyzed and extracted.Combining RDD with remote sensing image processing, Spark will record the trajectory of RDD transformation operation before trigger operation, the trajectory of RDD transformation will be automatically transformed into DAG graph, and after the trigger operation, the operation will be carried out according to the trajectory of DAG graph.And then the inversion index data modeling and processing calculation and the RDD conversion of the forest depression degree are carried out, and the parallel inversion of the forest canopy closure degree is completed.In this way, Spark avoids repeated operation and improves the computational efficiency of parallel inversion.

Numpy(Numerical Python) is a package that implements statistical functions which provides support for multidimensional arrays in Python. We selects the numpy statistical package in Python to select the relevant factors. First, run pyspark, to create a SparkConter object, and when using Python shell, the system automatically creates a SC variable. Spark uses the Cpython interpreter to support the use of the numpy statistics package. Transfer np.corrcoef(a,rowvar=0) to calculate the correlation coefficient between the parameters and the canopy density.

When calling numpy, use spark-submit to submit, call RDD:Self.rdd=self.sc.text.File("hdfs://server1:9000/user/w

TABLE 1. Correlation of various factors in GF1 image.

Variable factor	Correlation coefficien	Unilateral Sig.
NDVI	0.285	0.000
RVI	0.168	0.001
DVI	0.231	0.000
ARVI	0.156	0.000
X_{17}	0.197	0.000

ords.txt) The RDD converted from the interface of the generated wrapper class starts Spark-based distributed parallel computing based on the Spark map operation. Extract the top five of the correlation coefficients(That is, the significance of the influence on the dependent variable) from large to small. As shown in table 1:

In the table above, $X_{17} = \frac{\rho R \times \rho NIR}{\rho G}$.

Then, in the establishment of the inversion model, the Spark program is used to calculate the value of β_0, β_1, β_2, β_3 in formula (12). The preprocessed remote sensing files will be saved in HDFS in the form of images. Each block maintains a two-dimensional array. During the calculation, the map function is run every time a row of data is read, and the data in RDD is mapped to another RDD until the block is finished. The results of each calculation are key-value pairs to $<$ key, Value $>$ is stored in memory. The conversion operator provided by Spark is combined with the model to perform inversion operation, and the values with the same key are added together. The results of each operation are stored on the master node.The dependency relationship between RDD is realized by optimizing the algorithm and the parallel inversion method. The final inversion results are stored in HDFS.

Based on the above analysis, a remote sensing inversion model of forest canopy density in Laoshan experimental forest farm of Maoershan is established.

$$Y = 3.674 + 2.731\text{NDVI} - 0.263\text{RVI} + 0.027\text{DVI} - 4.321\text{ARVI} + 0.021\text{X}_{17} \quad (16)$$

Algorithm 1 Function np.corrcoef

```
input: Correlation factor and Canopy closure
output: sc.np.corrcorf(x.arrayfreq).map.
(x.calculate).collect;
sc=sparkcontext()
def correlation(x,y):stdev_x=standard_deviation(x)
stdev_y=standard_deviation(y)
if stdev_x > 0 and stdev_y > 0:
return covariance(x, y) / stdev_x / stdev_y
else:
return 0;
```

Algorithm 2 Function Multiple Linear Regression Model

```
input: read_input(sys.stdin)
output: Value of β0, β1, β2, β3, β4, β5
def matmulti()
if innerLength == 0:
innerLength = len(fields) − 1
data1 = np.array([0.0 for _ in range(innerLength)])
temp=np.array(fields,float)[:innerLength]*float
(fields[innerLength])
if length == 0:
length = len(fields) − 1
data2 = np.diag(np.zeros(length))
for index in range(length): data2[index]
= np.array(fields[:length]
data=data1+data2
```

VI. EXPERIMENT AND RESULT ANALYSIS

A. DATA PROCUREMENT AND REGIONAL SITUATION

Maoershan Laoshan experimental forest farm is located in the west of Zhangguangcailing Town, Shangzhi City, Heilongjiang Province which is an important experimental forest farm for teaching and scientific research of Northeast Forestry University. The forest is dense, the species is various, the life diversity is high.The Laoshan experimental forest farm has suffered a great damage in the early 20th century. After more than 50 years of restoration and reconstruction after the founding of the People's Republic of China, the area, volume, hectare volume and coverage of forests have increased year by year, and the structure has become more reasonable, which has made Laoshan Forest Farm gradually become a mixed forest ecosystem with mosaic distribution of artificial forests and natural forests. It has played a role in protecting the ecological environment of forests and maintaining the ecological functions of the region [32].

The paper takes GF-1 remote sensing image of Laoshan experimental forest farm in Maoershan as data source in July 2016, and its spatial resolution is 16m. Using ENVI to pretreatment remote sensing images, such as radiation correction, geometric correction and image enhancement, the error of geometric correction pixels is less than 0.5. The experimental data are the measured data of 80 sample points in Laoshan Experimental Forest Farm of Northeast Forestry University in 2016. Among them, 60 sample points are used to establish the inversion model and 20 sample points are used to test the accuracy of the model.

B. EXPERIMENTAL ENVIRONMENT

We use five Inspur Tidal INC I800 blade servers. (Xeon E5-2620V2, CPU6 cores, 12 threads, 2.10GHz frequency, 8G memory, 200GB SATA hard disk), one as the master node and four as the computing nodes, install Hadoop-2.6.5, Spark1.6.1, zookeeper-3.4.5 in these four nodes to build Hadoop clusters and Spark clusters in these four nodes. And use Tenda TEG1024G Gigabit Ethernet switch connection to run the cluster.

C. EVALUATION OF INVERSION ACCURACY

The inversion accuracy is carried out by three evaluation indexes. They are R-squared(R^2), Root Mean Square Error (RMSE) and Estiation Accuracy (EA). High R^2 and Low

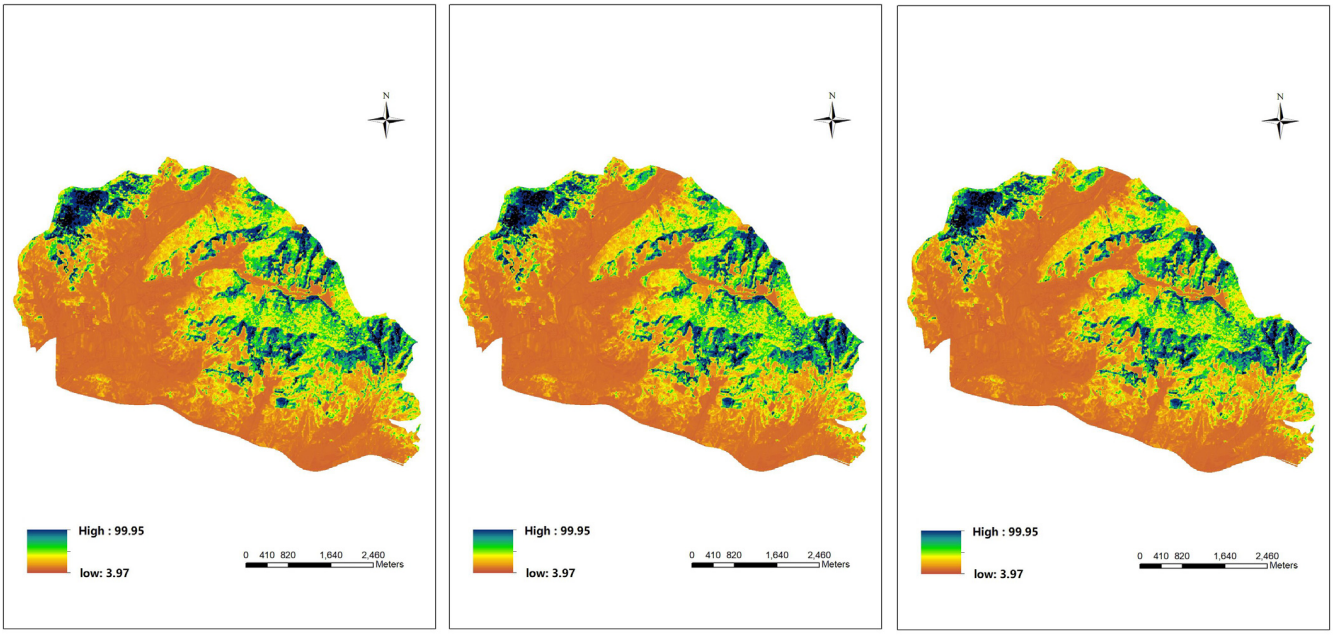

FIGURE 2. Parallel serial inversion and inversion results figure (a) is forest canopy inversion image with single computer serial, (b) is parallel inversion of image for forest canopy based on spark standalone, (c) is parallel inversion of image for forest canopy based on spark on yarn.

TABLE 2. Evaluation value of inversion accuracy.

R^2	RMSE	EA
0.651	0.0023	87.65

RMSE shows that the model has strong predictive power normally [33]. The value of R^2 above 0.4, low RMSE and the value of EA above 80 shows that the model has strong prediction ability and good inversion result in the inversion of canopy closure of forest.

The formula for calculating the index is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (yi - \hat{y}i)^2}{\sum_{i=1}^{n} (yi - \hat{y})^2} \tag{17}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (yi - \hat{y}i)^2}{n-1}} \tag{18}$$

$$EA = (1 - \frac{RMSE}{\text{mean}}) \times 100 \tag{19}$$

In above formula, yi is the measured sample of number i, $\hat{y}i$ is the estimated value of the corresponding i measured sample of number i. That is to say, the Y value calculated in equation (16) means the mean of the test sample data.

The results are shown in Table 2 below.

The results show that the value of R^2 is 0.651, above 0.4. The value of RMSE is small, and the value of EA is 87.65, higher than 80 which is a high precision. Therefore, the model can be used as the inversion model of canopy density of Laoshan experimental forest farm in Maoershan.

D. CONTRAST EXPERIMENT OF PARALLEL INVERSION OF CANOPY BASED ON SPARK

In this study, the accelerating effect of Spark on the data processing of remote sensing is verified by two groups of contrast experiments based on Spark standalone by mode and Spark on yarn mode. The parallel inversion model is evaluated from the accuracy of the forest canopy inversion and the efficiency of the inversion.

1) CONTRAST EXPERIMENT OF SPARK-STANDALONE PARALLEL MODEL AND SERIAL INVERSION

The standalone mode, which is a unique resource scheduling management mode in Spark, is the default cluster management mode implemented within Spark. One Master is responsible for managing the entire Cluster, without relying on any other resource management system. It is composed of Master and Worker nodes, and the cluster is mainly composed of Master and Worker nodes to achieve the typical Master/Slave mode. Application interactively applies for resources through Master nodes, and Worker nodes start Executor to run.

Experiment 1 is carried out in the distributed deployment mode of Spark-standalone, and the parallel inversion image of canopy density is obtained. In figure 2, (a) is the remote sensing image of forest canopy closure inversion by single computer, (b) is parallel inversion of remote sensing image for forest canopy-close degree based on Spark-standalone. Open-cv is used to compare the histogram of graph (b) and graph (a) to determine that the inversion results are consistent, and the inversion results are true and credible.

The experimental results are shown in Table 3: In the case of the same object being processed,the parallel inversion time based on Spark-standalone is 200749ms. The inversion time of single computer is 363065ms. The experimental results show that the parallel inversion rate is higher than that of the serial inversion, and the acceleration ratio is 1.8 times of that of the serial inversion.It is shown that the inversion efficiency based on Spark standalone is higher on the basis

TABLE 3. Inversion time and speedup in single machine modeand spark on yarn models.

Single-machine serial inversion time/ms	Parallel inversion time/ms	S_p
363065	200749	1.81

of the accuracy and reality of the parallel inversion results of forest depression density.

Speedup(S_p): The ratio of serial inversion on a single computer time to parallel inversion time, Where p represents p parallel nodes. In distributed system, acceleration ratio is one of the important indexes to measure the performance of distributed system.,which expressed as:

$$S_p = \frac{T_1}{T_P} \tag{20}$$

2) CONTRAST EXPERIMENT OF SPARK ON YARN AND SPARK-STANDALONE PARALLEL MODEL AND SERIAL INVERSION

Experiment 2 is carried out in the spark on yarn mode. Yarn is at the bottom of the entire data storage system and plays the most important role. It is responsible for resource scheduling of the entire system. Spark is responsible for scheduling and computing storage in each block RDD data set [34]. Apply for computing resources through Yarn, submit computing tasks and accept scheduling and integration of resources. The emergence of the Yarn resource manager has greatly improved the utilization of the cluster on the resource manager [35].

The parallel inversion image of canopy density is obtained. In figure 2, (a) is the remote sensing image of forest canopy closure inversion by single computer, (c) is parallel inversion of remote sensing image for forest canopy-close degree based on Spark on yarn. Open-cv is used to compare the histogram of graph (c) and graph (a) to determine that the inversion results are consistent, and the inversion results are true and credible.

The results of experiment 2 are shown in Table 4. Under the same experimental environment and data volume, the parallel inversion time based on Spark on yarn is 152083 ms, and the speedup ratio is 2.39, which is higher than that of experiment 1 (1.81). Therefore, the efficiency of parallel inversion based on Spark on yarn is higher than that based on Spark-standalone. In order to further explore the factor of efficiency of parallel inversion, we improve the parallel computing node to 9 based on Spark on yarn to carry out a comparative experiment. The experimental results are shown in figure 3. Under the same experimental environment and data, the time of parallel inversion decreases and the speedup of parallel inversion increases with the increase of computing nodes. When one node to two nodes, the growth rate is the fastest, reaching the peak rate.

VII. CONCLUSION

Two parallel methods of distributed deployment based on Spark-standalone and Spark on Yarn are proposed for the large amount of remote sensing data in forest crown closure inversion. Taking Maoershan Laoshan experimental forest farm of Northeast Forestry University as the experimental source, the related experiments are analyzed by using Spark parallel programming model and multiple linear regression model. The results show that: (1) The parallel inversion results based on the two models are consistent with the serial inversion results by comparing and analyzing the accuracy of the images, and the inversion results are accurate and reliable. (2) The rate of parallel inversion is significantly higher than that of serial inversion. The speedup ratio of parallel inversion based on Spark-standalone is 1.81 and the speedup ratio of parallel inversion based on Spark on yarn is 2.39 which is higher than based on Spark-standalone. At the same time, the acceleration ratio increases with the increase of computing node. When one node to two nodes, the growth rate is the fastest, reaching the peak rate.

Which can be drawn, parallel inversion of forest crown closure based on Spark has good acceleration ratio and inversion efficiency. It is feasible and suitable for efficient parallel inversion of massive remote sensing data in the era of large data remote sensing.

TABLE 4. Inversion time and speedup on different node numbers.

Node numbers	Parallel inversion time/ms	S_p
2	152083	2.39
3	133480	2.72
4	115995	3.13
5	100473	3.61
6	92563	3.92
7	87624	4.14
8	83569	4.34
9	78471	4.63

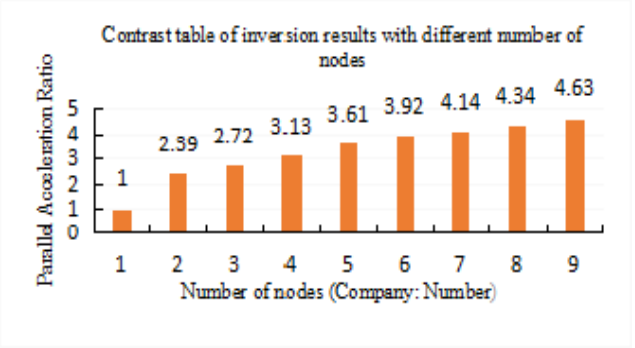

FIGURE 3. Contrast map of inversion results of different node numbers of forest canopy density.

REFERENCES

[1] T. Fu, Y. Pang, and Q. F. Huang, "Estimation of subtropical forest parameters by airborne LiDAR," *Remote Sens. J.*, vol. 15, no. 5, pp. 1092–1104, 2011.

[2] X. Y. Mu, Q. L. Zhang, and Q. W. Liu, "Inversion of stand average height and canopy density based on airborne LiDAR data," *J. Northeast Forestry Univ.*, vol. 43, no. 9, pp. 84–89, 2015.

[3] F. R. Yu, J. Gao, and J. Fu, "Research on extraction method of forest canopy density based on Tencent Street view," *J. Southwest Forestry Univ., Natural Sci.*, vol. 38, no. 5, pp. 139–144, 2018.

[4] C. G. Li and T. J. Cai, "The influence of forest canopy on the estimation of stocking capacity," *J. Northeast Forestry Univ.*, vol. 34, no. 1, pp. 15–17, 2006.
[5] Y. N. Li, B. L. Zhang, and S. Y. Qin, "Study on canopy closure and its determination method and its application," *World Forestry Res.*, vol. 21, no. 5, pp. 40–46, 2008.
[6] Y. Li, C. Xu, Y. Teng, X. Cheng, and B. Yu, "Determination of canopy density of larch forest in North China by digital photographs," *J. Northeast Forestry Univ.*, vol. 38, no. 11, pp. 34–37,2010.
[7] B. Xu, P. Gong, and R. Pu, "Crown closure estimation of oak savannah in a dry season with Landsat TM imagery: Comparison of various indices through correlation analysis," *Int. J. Remote Sens.*, vol. 24, no. 9, pp. 1811–1822, 2003.
[8] B. X. Tan, Z. Y. Li, and X. E. Chen, "Quantitative estimation of forest canopy with Hyperion hyperspectral data," *J. Beijing Forestry Univ.*, vol. 28, no. 3, pp. 95–101, 2006.
[9] N. Wang, "Dynamic monitoring of forest resource change based on 3S technology," Ph.D. dissertation, Nanjing Forestry Univ., Nanjing, China, 2012.
[10] I. A. Iqbal, J. Dash, S. Ullah, and G. Ahmad, "A novel approach to estimate canopy height using ICESat/GLAS data: A case study in the New Forest National Park, UK," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 23, pp. 109–118, Aug. 2013.
[11] D. Moeser, J. Roubinek, P. Schleppi, F. Morsdorf, and T. Jonas, "Canopy closure, LAI and radiation transfer from airborne LiDAR synthetic images," *Agricult. Forest Meteorol.*, vol. 197, no. 19, pp. 158–168, Oct. 2014.
[12] S. L. Wu, *Forest Canopy Inversion Based on Landsat 8 OLI Data*. Beijing, China: Beijing Forestry Univ., 2014.
[13] X. Cai and Y. Yang, "Estimation of forest canopy density based on CMB satellite data," *Anhui Agricult. Sci.*, vol. 35, no. 34, pp. 10961–10962, 2007.
[14] R. Y. Zhang, Y. Pang, and Z. Y. Li, "Estimation of temperate forest canopy with airborne LiDAR and LANDSAT ETM data," *J. Plant Ecol.*, vol. 40, no. 2, pp. 102–115, 2016.
[15] Z. Y. Niu, J. Feng, and J. C. Gu, "Remote sensing inversion of forest canopy based on leaf area index," *Forestry Resource Manage.*, vol. 1, pp. 46–51, 2014.
[16] R. L. Pu, B. Xu, and P. Gong, "Oakwood crown closure estimation by unmixing Landsat TM data," *Int. J. Remote Sens.*, vol. 24, no. 22, pp. 4433–4445, 2003.
[17] R. L. Pu, B. Xu, and P. Gong, "Wavelet transform applied to EO-1 hyperspectral data for forest LAI and crown closure mapping," *Remote Sens. Environ.*, vol. 91, no. 2, pp. 212–224, 2003.
[18] Q. Liu, L. Yang, and Q. H. Liu, "Summary of remote sensing inversion methods for forest aboveground biomass," *Remote Sens. J.*, vol. 19, no. 1, pp. 62–74, 2015.
[19] R. E. Bellman, *Adaptive Control Process*. Princeton, NJ, USA: Princeton Univ. Press, 2015.
[20] F. Le, L. Wang, J. Liu, and Q. Chang, "Remote sensing estimation of winter wheat leaf SPAD value based on Gaofen-1 satellite data," *J. Agricult. Mach.*, vol. 46, no. 9, pp. 273–281, 2015.
[21] D. Li, "On the development of remote sensing and GIS in the 21st century," *J. Wuhan Univ.*, vol. 25, no. 2, pp. 127–131, 2003.
[22] H. Meng, "Research on normalized vegetation index algorithm of remote sensing image based on GPU," Ph.D. dissertation, Henan Univ., Kaifeng, China, 2016.
[23] T. X. Fu, Z. J. Liu, and H. W. Yan, "Research on biomass remote sensing parallel inversion method based on MapReduce model," *Resour. Environ. Arid Areas*, vol. 271, no. 1, pp. 130–136, 2013.
[24] Z. J. Li, X. J. Li, and T. Liu, "Review of MapReduce programming model and its application in image processing," *Mapping Spatial Geograph. Inf.*, vol. 38, no. 4, pp. 19–21, 2015.
[25] G. Ning *et al.*, "Fast construction method of tile pyramid for large scale raster data set," *Geograph. Inf. World*, vol. 22, no. 6, pp. 43–50, 2015.
[26] C. Xia, J. Li, and Q. Liu, "Review of advances in vegetation phenology monitoring by remote sensing," *J. Remote Sens.*, vol. 17, no. 1, pp. 1–16, 2013.
[27] G. Huang and J. Guo, "Data division in distributed parallel remote sensing image processing," *Remote Sens. Inf.*, vol. 2, pp. 9–12, 2001.
[28] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop distributed file system," in *Proc. IEEE 26th Symp. Mass Storage Syst. Technol. (MSTT)*, Incline Village, NV, USA, May 2010, pp. 1–10.
[29] A.-K. Seghouane, "New AIC corrected variants for multivariate linear regression model selection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 47, no. 2, pp. 1154–1165, Apr. 2011.
[30] J. F. Leng, X. Gao, and J. Zhu, "Application of multiple linear regression statistical prediction model," *Statist. Decision Making*, vol. 7, pp. 82–85, 2016.
[31] Y. J. Kaufman and D. Tanre, "Atmospherically resistant vegetation index (ARVI) for EOS-MODIS," *IEEE Trans. Geosci. Remote Sens.*, vol. 30, no. 2, pp. 261–270, Mar. 1992.
[32] Z. Ding, Y. J. Yao, and R. Q. Song, "Historical changes of Forest Resources in Laoshan area of Maoershan experimental forest farm," *J. Northeast Forestry Univ.*, vol. 34, no. 3, pp. 87–89, 2006.
[33] H. Du, W. Fan, G. Zhou, X. Xu, H. Ge, Y. Shi, Y. Zhou, R. Cui, and Y. Lu, "Retrieval of canopy closure and LAI of Moso bamboo forest using spectral mixture analysis based on real scenario simulation," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 11, pp. 4328–4340, Nov. 2011.
[34] J.-M. Morel, A. B. Petro, and C. Sbert, "Fourier implementation of Poisson image editing," *Pattern Recognit. Lett.*, vol. 33, no. 3, pp. 342–348, Feb. 2012.
[35] J. Xia, *Spark Big Data Processing Technology*. Beijing, China: Electronic Industry Press, 2015.

GUANGSHENG CHEN received the Ph.D. degree. He is a doctoral supervisor. He has published more than 30 academic articles and one monograph. He has presided over and participated in more than ten projects such as the National Science and Technology Support Project, the National Science and Technology Foundation Project, the National Natural Science Fund, the National Public Welfare Industry Research Project, the Heilongjiang Science and Technology Tackling Plan, and the Provincial Natural Science Fund. He has won more than three second-class prizes of provincial and ministerial self-government for natural science and technological progress.

TONGTONG LOU is currently pursuing the master's degree with Northeast Forestry University, China. Her current research interests include cloud computing, modern information technology, and digital forestry.

WEIPENG JING received the Ph.D. degree from the Harbin Institute of Technology of China. He is currently an Associate Professor with Northeast Forestry University, China. His research interests include modelling and scheduling for distributed computing systems, fault tolerant computing and system reliability, cloud computing, and spatial data mining. He has published more than 50 research articles in refereed journals and conference proceedings such as CPC, PUC, and FGCS.

ZEYU WANG received the master's degree from the School of Information and Computer Engineering, Northeast Forestry University of China. His current research interests include big data cloud computing, modern information technology, and artificial intelligence.