SUPPLEMENTARY MATERIAL

S1. Comparison between the GF-2, Landsat 8 and Fusion Images

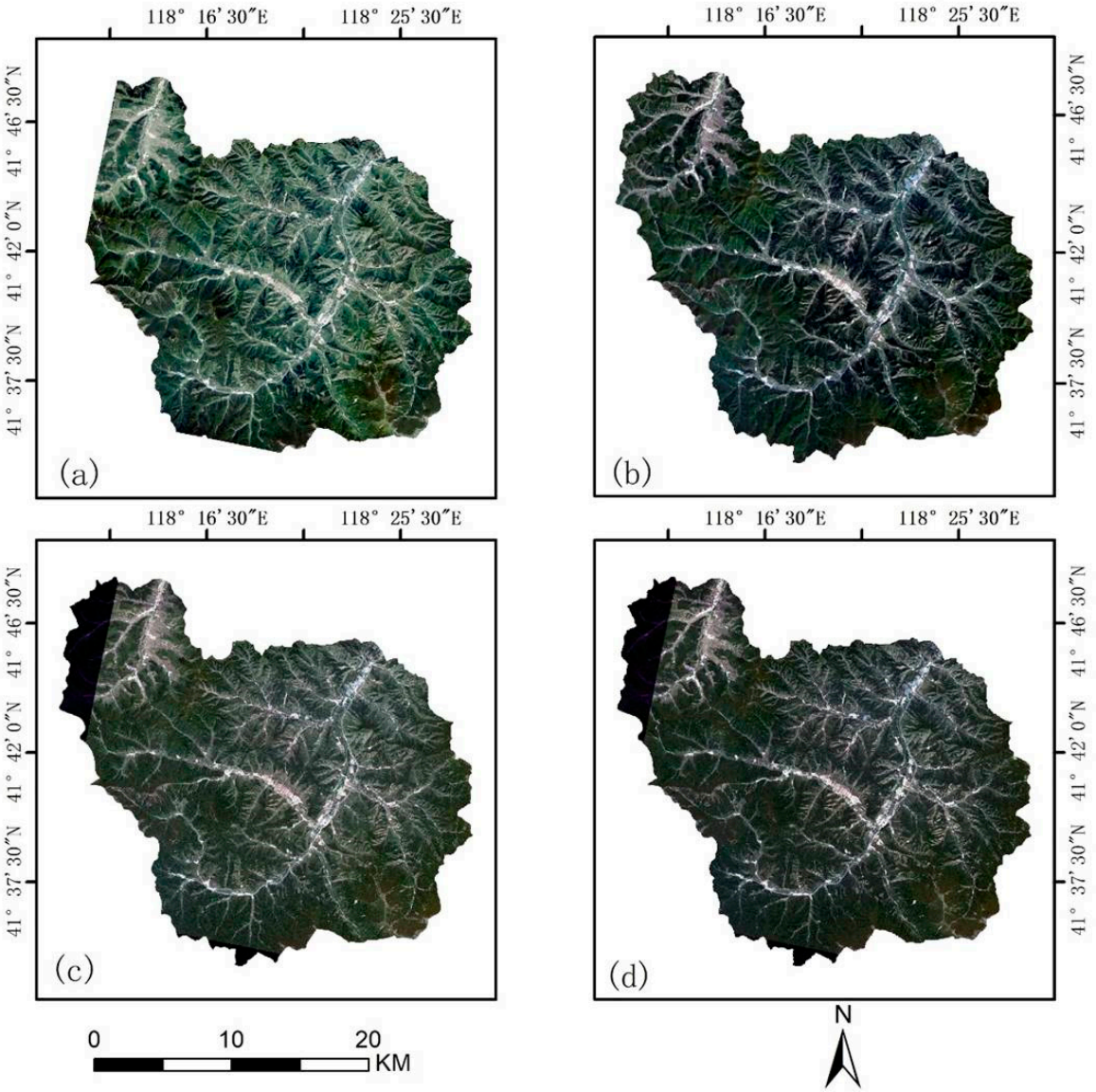

Figure S1. (a) The GF-2 image; (**b**) the Landsat 8 image; (**c**) the fusion image Green_Landsat; and (**d**) the fusion image Red_Landsat.

S2. Selection of Suitable Variables using DC-FSCK

In this study, a DC-FSCK approach that integrated feature variable screening and a combination optimization procedure based on distance correlation coefficient and k-nearest neighbors (kNN) algorithm was proposed and compared with stepwise regression analysis (SRA) and random forest (RF) for feature variable selection. The flow chart of DC-FSCK is shown in Figure S2.

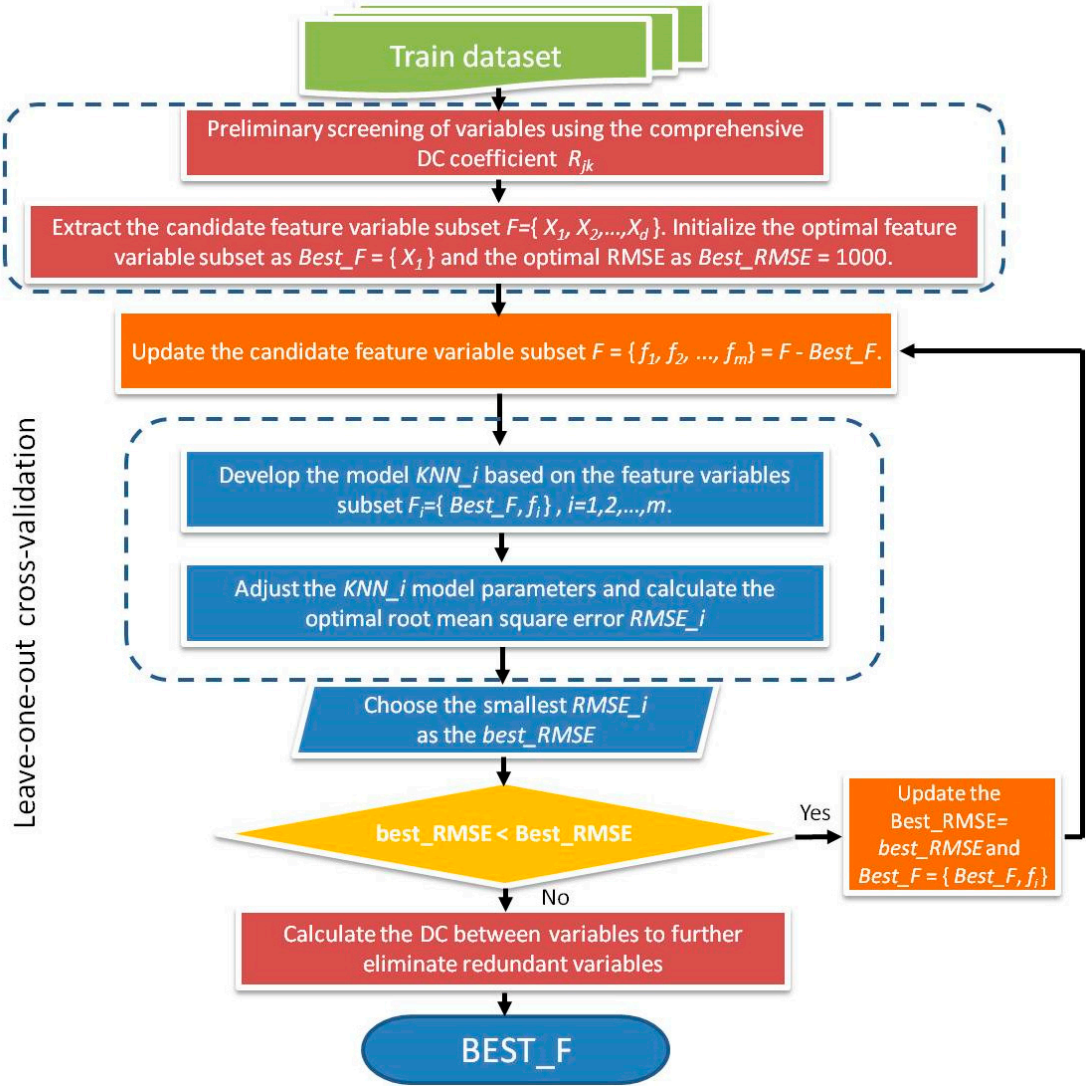

Figure S2. The process of selecting feature variables by the DC-FSCK algorithm.

S3. Stepwise Regression Analysis and Random Forest for Feature Variable Selection

In order to assess the superiority of DC-FSCK in GSV estimation, we compared it with Stepwise Regression Analysis (SRA) and Random Forest (RF) for spectral variable selection.

The SRA is a commonly used method for screening feature variables [28,30,31]. Since the number of the spectral variables is larger than the number of the sample plots, we first explored the relationship between the feature variables and the GSV and the correlations among the spectral variables. Only the variables that were strongly related to the GSV but weakly correlated with other feature variables were selected. The RF can optimize selection of feature variable by comparing the estimation error before and after a feature variable is removed [19,49]. A great increase of error means that the variable is important and otherwise, it is not. All the variables are ranked according to the importance. Then, a Pearson correlation analysis was performed on the selected variables [19]. If two variables have a large correlation coefficient, the less important one was removed. The results of model estimation using the selected spectral variables are evaluated based on the adjusted determination coefficient R_2 and relative root mean square error [31].

S4. Development of GSV Estimation Models

In this study, we compared two machine learning regression algorithms (kNN: k-nearest neighbors, SVR: Support Vector Regression), three ensemble learning algorithms (RF, XGBoost: eXtreme Gradient Boosting, and Stacking) and a parametric approach, Multiple Linear Regression (MLR).

The MLR was used to estimate GSV with the assumption that the feature variables from the remotely sensed images have a linear relationship with the field plot GSV [28]. Thus, selecting suitable feature variables for the model was important.

The kNN is one of the simplest machine learning algorithms [65], which needs to a set of input parameters, including the type of spectral distance metric, weighting function, and k value. In this study, the Minkowski distance was used. The weight was a function inversely proportional to the spectral distance indicating similarity or dissimilarity, and the sum of the weights of k nearest neighbors was set up to 1. The SVR is a statistical learning approach [52] and is an important application of the support vector machine (SVM). Using a small number of training samples, SVM can provide higher classification or estimation accuracy than other approaches. The SVR uses a nonlinear kernel function to minimize training errors and the model complexity by transforming input data into a high-dimensional feature space. Kernel based SVR methods have been commonly used, but some of their parameters need to be tuned, such as the kernel, precision and penalty parameters. The RF is a classification or regression estimation algorithm [19] that is used to estimate forest GSV on the basis of the results of multiple regression trees. If a large proportion of data is missing, RF can estimate the missing data and maintain the GSV estimation accuracy. The RF has two most important parameters, which are the number of regression trees and the number of the randomly selected variables. The XGBoost is a gradient Boosting-based emerging efficient integrated learning algorithm [54]. By adding a regular term to the cost function to minimize the complexity of the model, and taking the idea of RF, XGBoost enables random sampling of independent variables, which can reduce over fitting and simplify calculation.

The Stacking ensemble learning framework first divides the original dataset into several subsets and puts them into each base learners of the first layer prediction model. Each base learner outputs prediction results. The Stacking uses all the results of the first layer prediction model as the inputs of the second layer to train the meta-learners of the second layer prediction model and outputs the final prediction results. The Stacking generalizes the output results of multiple models to improve the overall prediction accuracy (Fig. S3) [57]. It is mostly used for classification research [56] and rarely utilized for the GSV estimation.

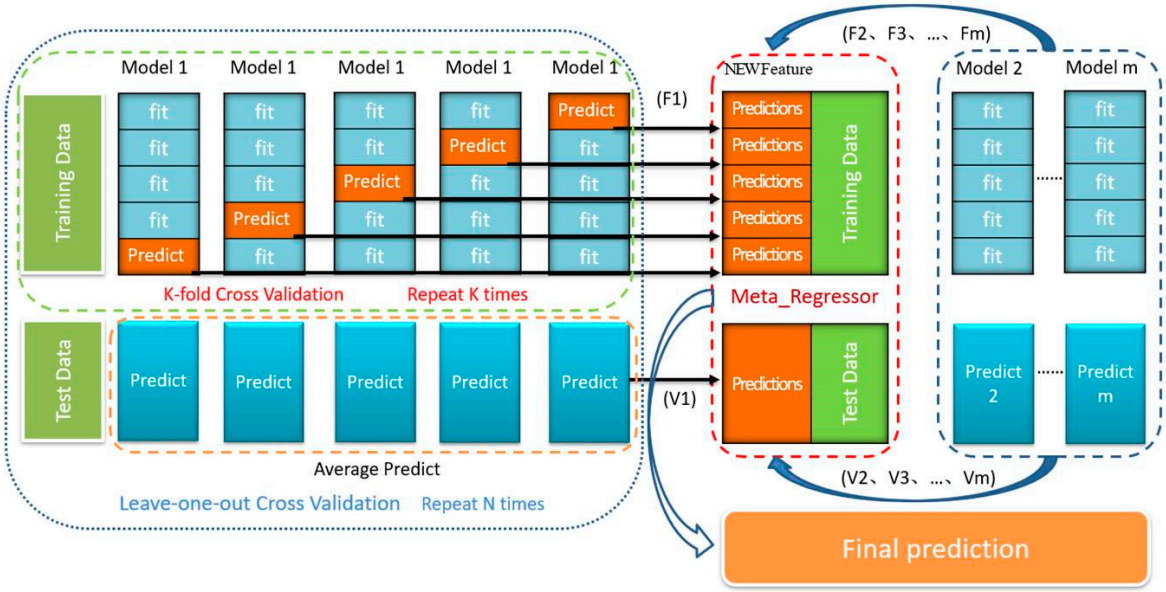

Figure S3. Framework of the Stacking algorithm