Simulation of the vegetation distribution in the Jing-Jin-Ji region that has been highly disturbed by social-economic development

Sangui Yi [1, 2], **Jihua Zhou** [1], **Liming Lai** [1], **Hui Du** [1], **Qinglin Sun** [1, 2], **Liu Yang** [1, 2], **Xin Liu** [1, 2], **Benben Liu** [1, 2], **Yuanrun Zheng** [Corresp. 1]

[1] Key Laboratory of Resource Plants, West China Subalpine Botanical Garden, Institute of Botany, Chinese Academy of Sciences, Beijing, China

[2] University of Chinese Academy of Sciences, Beijing, China

Corresponding Author: Yuanrun Zheng
Email address: zhengyr@ibcas.ac.cn

Background. Vegetation distribution simulations could help to understand vegetation distribution patterns and trends, but it is difficult to accurately simulate the distribution of vegetation especially in regions that are heavily affected by human disturbance.

Methods. Climate, topographic, and spectral data were used as input predictor variables of four machine learning models, including the random forest (RF), decision tree (DT), support vector machine (SVM) and maximum likelihood methods, in three vegetation classification units, including the vegetation group, vegetation type, and formation and subformation, in the Jing-Jin-Ji region, which is one of the most developed regions in China. A total of 2789 vegetation points were used for model training, and 974 vegetation points were used for model assessment.

Results. The result showed that the random forest method was the best of the four models and could simulate the distribution of the vegetation in all three classification units well. Kappa coefficients indicated that the random forest method had the highest prediction ability in regard to vegetation type, followed by vegetation group, formation and subformation. Five predictor variables, including 4 climate variables (annual mean temperature, max temperature of warmest month, min temperature of coldest month and annual precipitation) and 1 geospatial variable (elevation), were the most important for three vegetation classification levels. The winter surface albedo of band 4, the slope and the three summer spectral variables (the summer surface albedo of bands 2 and 6 and the summer brightness index) could also increase the accuracy of vegetation classification to some extent.

Conclusions. In all three levels, RF models performed well, while other three models could not simulate the distribution. The RF model was the best model for simulating the vegetation distribution in the Jing-Jin-Ji region. Four climate variables and one geospatial variable enhanced greatly the accuracy of vegetation classification, and the winter surface albedo of band 4, the slope, and the three summer spectral variables could also increase the accuracy of vegetation classification to some extent.

Simulation of the vegetation distribution in the Jing-Jin-Ji region that has been highly disturbed by social-economic development

Sangui Yi[1, 2], Jihua Zhou[1], Liming Lai[1], Hui Du[1], Qinglin Sun[1, 2], Liu Yang[1, 2], Xin Liu[1, 2], Benben Liu[1, 2], Yuanrun Zheng[1, *]

[1] Key Laboratory of Resource Plants, West China Subalpine Botanical Garden, Institute of Botany, Chinese Academy of Sciences, Xiangshan, Beijing, China.
[2] University of Chinese Academy of Sciences, Beijing, China.

Corresponding Author:
Yuanrun Zheng[1, *]
No. 20 Nanxincun, Xiangshan, Beijing, 100093, China
Email address: zhengyr@ibcas.ac.cn

Abstract

Background. Vegetation distribution simulations could help to understand vegetation distribution patterns and trends, but it is difficult to accurately simulate the distribution of vegetation especially in regions that are heavily affected by human disturbance.
Methods. Climate, topographic, and spectral data were used as input predictor variables of four machine learning models, including the random forest (RF), decision tree (DT), support vector machine (SVM) and maximum likelihood methods, in three vegetation classification units, including the vegetation group, vegetation type, and formation and subformation, in the Jing-Jin-Ji region, which is one of the most developed regions in China. A total of 2789 vegetation points were used for model training, and 974 vegetation points were used for model assessment.
Results. The result showed that the random forest method was the best of the four models and could simulate the distribution of the vegetation in all three classification units well. Kappa coefficients indicated that the random forest method had the highest prediction ability in regard to vegetation type, followed by vegetation group, formation and subformation. Five predictor variables, including 4 climate variables (annual mean temperature, max temperature of warmest month, min temperature of coldest month and annual precipitation) and 1 geospatial variable (elevation), were the most important for three vegetation classification levels. The winter surface albedo of band 4, the slope and the three summer spectral variables (the summer surface albedo of bands 2 and 6 and the summer brightness index) could also increase the accuracy of vegetation classification to some extent.

Conclusions. In all three levels, RF models performed well, while other three models could not simulate the distribution. The RF model was the best model for simulating the vegetation distribution in the Jing-Jin-Ji region. Four climate variables and one geospatial variable enhanced greatly the accuracy of vegetation classification, and the winter surface albedo of band 4, the slope, and the three summer spectral variables could also increase the accuracy of vegetation classification to some extent.

Introduction

Vegetation is an important resource that can provide many service functions for organisms because it is an important component of terrestrial ecosystems and landscapes (Editorial Committee of Vegetation Map of China, the Chinese Academy of Science, 2007). Environmental research, and natural resource management, and even urban or rural planning require maps of the distribution of vegetation or maps of vegetation habitat suitability, such as the assessment of biodiversity, the management, conservation and restoration of habitat, and even forecasting the impact of environmental changes on vegetation and ecosystems (Franklin, 2010). To better understand and smartly use vegetation, monitoring and mapping are highly necessary and essential. The pattern and distribution of the vegetation are affected by both climate and disturbances (Chen et al., 2015; Zhang et al., 2018), especially disturbances caused by land use change (Hansen et al., 2013; Wehkamp et al., 2018). Since industrialization, humans have strongly influenced the environment and vegetation, the vegetation pattern has been greatly changed, and the exact mapping of vegetation under fast and great disturbance has recently become a difficult challenge (Xie, Sha & Yu, 2008; Zhou et al., 2016).

Traditionally, field surveys are the main method to map vegetation; however, field surveys require labor and money (Newell & Leathwick, 2005; Zhou et al., 2016). Vegetation mapping using remote sensing data has become a popular method in the past 30 years. People can obtain a wide range of reliable data from remote sensing images. In addition, vegetation boundaries can be determined and updated by the visual interpretation of images and field surveys, respectively. However, it is quite unreliable to determine vegetation units and their boundaries by visual interpretation. Researchers may get different results even with the same images for the same study area because of their subjectivity (Bie & Beckett, 1973; Pfeffer, Pebesma & Burrough, 2003). Further, for methods based on field surveys and remote sensing, the boundaries of different vegetation units are manually drawn using information based on information, such as the climate, elevation, and soil type, which can cause inaccuracies in transition areas (Zhang et al., 2008). Using field data and remote sensing data with simulation models may be an alternative for mapping vegetation.

The environment where vegetation grows generally affects the composition, structure, and function of vegetation communities, which in turn affects the spatial distribution of vegetation. Environmental variables have been used to simulate the distribution of vegetation around the world (Dilts et al., 2015; Mod et al., 2016), and these models are usually developed from assumptions how environmental variables control the distribution of vegetation (Guisan &

Zimmermann, 2000). Therefore, the easy access to data such as terrain, soil and climate, and the use of geographic information system software for manipulating these data make it possible to computerized predictive vegetation map (Franklin, 1995).

Predictive vegetation mapping is developed based on niche theory and gradient analysis and driven by environmental research, natural resource management and even urban or rural planning (Franklin, 1995; Dilts et al., 2015). Actually, it projects the processes of community assembly onto geographic space using environmental variables through various models (Franklin, 1995; Lany et al., 2019), which is suitable to mapping the vegetation of a large landscape and analyzing the relationship between vegetation and the environment. Some methods based on statistics and machine learning have been used to simulate vegetation distribution. Predictive vegetation mapping includes various statistical methods such as the generalized linear model, the generalized additive model, multivariate statistical approaches and so on (Franklin, 2010). Recently, machine learning modeling methods have been used to map both the distribution of vegetation communities and individual species; these methods include support vector machines, decision tree sand artificial neural network (Guisan & Zimmermann, 2000; Hastie, Tibshirani & Friedman, 2009; Zhou et al., 2016). These machine learning models have fewer limitations and can produce more reliable results than traditional vegetation modeling methods (Hastie, Tibshirani & Friedman, 2009). Advanced machine learning techniques can integrate spectral and spatial predictors and retain important information about the vegetation composition and structural differences to improve the classification accuracy (Sesnie et al., 2010). With the development of remote sensing technology, large-scale high-resolution remote sensing images can be acquired in a short time. Machine learning models combined with other environmental data, including remote sensing data, can reduce the amount of field surveys and the visual interpretation of remote sensing images, which saves economic and labor costs and avoids inaccuracies caused by visual interpretation to a certain extent. Therefore, these models have become an effective method.

The Jing-Jin-Ji urban agglomeration, also known as the Beijing-Tianjin-Hebei urban agglomeration, is the center of Chinese politics and culture and an important core area of the northern Chinese economy. However, it also faces problems including unbalanced regional development and conflicts between economic development and limited resources. On the one hand, big cities, such as Beijing and Tianjin, have large populations, developed economies, and abundant educational resources. On the other hand, these big cities are facing problems of limited natural resources and serious ecological and environmental pollution. In particular, there is a big conflict between a huge population and limited resources, such as water, land and even vegetation, in Beijing (Wang & Gong, 2018). Therefore, for coordinated regional development, it is best to break administrative divisions and study the entire region (Wang et al., 2019). To evacuate the population of Beijing, the new Xiong'an area located in Hebei province has been planned and is being constructed. The development of these cities is affected by the surrounding natural environment. To better integrate the environmental carrying capacity and socioeconomic development for the Jing-Jin-Ji region, including the new Xiong'an area, accurate vegetation

maps with temporal resolution are necessary. The most recent vegetation map in the Jing-Jin-Ji region is the Vegetation Map of the People's Republic of China (VMC), with a scale of 1:1000000 (Editorial Committee of Vegetation Map of China, the Chinese Academy of Science, 2007), and most data are from a field survey from around 1980; both need to be updated in terms of their temporal and space scales.

In this paper, geospatial, climate, and spectral data were integrated to simulate vegetation distribution through different four models in different three vegetation classification levels. The purposes were to (1) determine the predictive ability of different vegetation models in different vegetation classification levels, (2) explore a suitable modeling method for vegetation in the Jing-Jin-Ji region affected by high social- economic disturbance, and (3) determine which predictive variables enhanced the accuracy of classification for vegetation mapping.

Materials & Methods

Study area

The Jing-Jin-Ji region is located in the northern area of the North China Plain, which ranges from 113° 04′ to 119° 53′ E and 36° 01′ to 42° 37′ N, and surrounded by Taihang Mountain in the west, Yanshan Mountain in the north and Bohai Sea in the east, including Beijing, Tianjin and Hebei Province (Fig. 1). It has a population of approximately 110 million and covers an area of approximately 216,000 km^2 (Wang et al., 2019). Temperate monsoon climate zone covers this region. The elevation ranges from -14 to 2837 m (Fig. 1). The annual precipitation ranges from 305 to 711 mm, with higher precipitation amounts at lower altitudes. The annual mean temperature ranges from -3 to 14 °C, with cooler averages at higher elevations. The precipitation amount gradually decreases from the southeast to the northwest in the study area, but the temperature shows the reverse pattern.

Vegetation and training data

The VMC, as the most recent vegetation map, contains 8 vegetation groups, 15 vegetation types, and 75 formations and subformations in the Jing-Jin-Ji region. The VMC was completed in 2007 based on field survey data. However, the areas of some vegetation units are too small, making them very difficult to distinguish. Therefore, we selected 8 vegetation groups, 12 vegetation types, and 39 formations and subformations in the study area (Table 1). Cultivated vegetation areas are mainly distributed in low areas where the altitude mainly ranges from -14 to 254 m and the annual mean temperature mainly ranges from 7 to 14℃ and mainly consist of winter wheat and coarse grains. Scrub and grass-forb communities are mainly distributed in the north, ranging from 254 to 1440 m.

The model training and assessment data were obtained from field surveys and publications. These data contained information on vegetation compositions. A total of 3763 vegetation points were collected, of which 2789 were used for model training and 974 were used for model assessment. There are at least 80 vegetation points for each formation and subformation, of which at least 60 vegetation points were used for model training and 20 vegetation points were

used for model assessment. In addition to vegetation point data, the VMC was also used for model assessment to increase the credibility of the model assessment.

Geospatial, climate, and spectral data

Geospatial variables, including elevation, slope, and aspect, were derived from the 30-m resolution SRTM DEM product (Zhao et al., 2018). These data were resampled to a 500×500 m grid cell size using a nearest-neighbor method in ArcGis 10.3 (Chang, Chen & Lian, 2014).

Climate data at a 1-km resolution included the annual mean temperature, max temperature of warmest month, min temperature of coldest month and annual precipitation downloaded from WorldClim (Fick & Hijmans, 2017) at http://worldclim.org/. These climate data were also resampled to a 500×500 m grid cell size using a nearest-neighbor method in ArcGIS 10.3 (Chang, Chen & Lian, 2014). These climatic variables are ecophysiologically meaningful variables for plants (Mod et al., 2016) and are commonly used as bioclimatic limits in vegetation models (Sitch et al., 2003; Franklin, 2010).

The MYD09A1500M product data (sinusoidal projection, path 4 and row 26, path 4 and row 27, path 5 and row 26, path 5 and row 27) in summer (July 20, 2013) and winter (January 17, 2013), as Modis images, were acquired from the Geospatial data Cloud at http://www.gscloud.cn/. Image pre-processing included image subset mosaicking and image clipping according to study area in ENVI 5.2 (Deng, 2010). The land surface albedo in bands 1-7 was directly obtained from the MYD09A1500M product, and vegetation indices that were proven effective to reflect vegetation information (Price, Guo & Stiles, 2002; Zhou et al., 2016) were calculated.

Vegetation indices indicate the mixtures information of vegetation and its surrounding including soil, light, moisture and so on (Bannari, Morin & Bonn, 1995), which is more sensitive than a single spectral band for green vegetation detection (Bannari, Morin & Bonn, 1995; Cohen & Goward, 2004). Therefore, vegetation indices can be used for images interpretation, the discrimination and prediction of different vegetation, the evaluation of vegetative cover density, and even the detection of land use changes (Bannari, Morin & Bonn, 1995; Goward, 2004; Zhou et al., 2016).

A total of 49 variables, including 3 geospatial variables, 4 climate variables and 42 spectral variables (7 land surface albedos and 14 vegetation indices for summer and winter) were used for model training and assessment (Table 2). Through different model methods, we used different combinations of variables to simulate the distribution of vegetation. Combination 1-6 contained at least one seasonal spectral variable, either land surface albedos or vegetation indices. Combination 7-9 contained at least one geospatial variable or climate variable. Combination 10 contained all 49 variables, including all spectral, climate and geospatial variables. Combination 11 contained the top 10 most important variables of the decision tree, with all variables in the vegetation group level (DT10). Combination 12 contained the top 10 most important variables of the random forest method, with all variables in the vegetation group level (RF10). The support vector machine and the maximum likelihood classification (MLC) methods only output the

simulation results of variable combinations 1-6, likely due to the poor separability of the training samples (Deng, 2010).

Vegetation distribution models

Decision tree, random forest, maximum likelihood classification and support vector machine models were used to model the vegetation distribution in our study. DT models are divisive, monothetic, supervised classifiers, and many people use them for species distribution modeling and related applications (Franklin, 2010). The DT model is computationally fast and easy to understand and implement. It generates classification rules through classification or regression algorithms, and then visualize the classification rules into simple tree graphics (Hastie, Tibshirani & Friedman, 2009; Zhang & Dong, 2013; Zhou et al., 2016). And the DT model calculates the important variables that contribute greatly to the model (Deng, 2010). The classification rules of DT models in this study were with 5 layers, where 40 samples were in the smallest parent node and 10 samples were in the smallest child node.

The RF model is an ensemble method and has been applied in some risk assessment and species distribution modeling studies (Cutler et al., 2007; Franklin, 2010; Zhang & Dong, 2017). RF models fit many classification trees to a data set, the diversity of which is ensured by the use of random samples derived from the training dataset, and then combine the predictions from all the trees to produce considerably more accurate classifications by combining many classification trees; they are not sensitive to noise or overtraining (Gislason, Benediktsson & Sveinsson, 2006; Cutler et al., 2007; Burai et al., 2015). And the RF model also calculates the important variables that contribute greatly to the model (Cutler et al., 2007). Classification rules of the RF models in this study were with the default settings in EnMAP-Box, where 100 trees existed and the node impurity function was with Gini coefficient (van der Linden et al., 2015; Zhou et al., 2016).

The MLC model is one of the most commonly used supervised image classification methods. The classification rules of MLC assign every pixel to the class with the highest probability according to the statistics of the Gaussian probability density function. MLC method is not applicable when there are fewer training samples and more input predictors, even it usually generates similar or more accurate classification than other methods (Burai et al., 2015; Zhou et al., 2016).

The SVM model is a supervised machine learning model that can be used for classification and regression. The SVM model is a complex and widely used method that can output more accurate predictions (Burai et al., 2015). The SVM model aims to find an optimal plane in a multidimensional space that divides all sample elements into two categories, which should make the distance between the closest points in the two classes as large as possible (Kabacoff RI., 2016). Similarly, the default settings in EnMAP-Box are also applied to the SVM model in this study. (van der Linden et al., 2015).

The predicted vegetation maps of three classification units, including vegetation groups, vegetation types and formations and subformations, were generated through DT, RF, MLC and SVM methods. Their resolution was 500 m. We selected all twelve variable combinations as

input variables for every method. Important variables were generated by the DT and RF methods, and they were important for the simulation of vegetation distribution.

Model assessment

A total of 974 vegetation points and the VMC were used to assess all predicted vegetation maps, which generated the overall accuracy and Kappa coefficient of every predicted vegetation map. It is generally considered that when the value of the Kappa coefficient is greater than 0.4, the assessed predicted map is acceptable. The Kappa coefficient value is generally defined to range from 0.4 to 0.55 to indicate moderate agreement, from 0.56 to 0.8 to indicate substantial agreement and from 0.81 to 1 to indicate almost perfect agreement (Landis & Koch, 1977; Weng & Zhou, 2006; Zhou et al., 2016).

Results

Vegetation group modeling and assessment

Among DT, RF, MLC and SVM models, the results of RF models were better than the results of the other three models. Only RF model had the Kappa coefficient larger than 0.4 using variable combinations 3 and 6-12 assessed by field point data, with overall accuracy were from 50% to 72%; RF model had the Kappa coefficient larger than 0.56 using variable combinations 8-12 assessed by field data with overall accuracy were from 67% to 72%; the highest Kappa coefficient of 0.67 with the highest overall accuracy of 72% existed in the RF models with variable combination 12 (Table 3). When assessed by the VMC, the highest Kappa coefficient of 0.38 with overall accuracy of 57% existed in the RF models with the variable combinations 10 (Table 3, Fig. 2).

Vegetation type modeling and assessment

Same as vegetation groups modeling, the results of RF models were better than the results of the other three models. In the RF models assessed by field point data and the VMC, the results using variable combinations 8-12 had the Kappa coefficient lager than 0.4 with overall accuracy of 65%-71% and 53%-55%, respectively. The Kappa coefficient larger than 0.56 existed in the RF models with overall accuracy of 65%-71% using variable combinations 8-12 assessed by field point data. For assessment by field point data, the highest Kappa coefficient of 0.67 and the highest overall accuracy of 71% existed in the RF models with variable combination 9; for assessment by the VMC, the two best results existed in the RF model using variable combination 9-10, the Kappa coefficient were both 0.43 with same overall accuracy of 55% (Table 4, Fig. 3).

Formation and subformation modeling and assessment

The results of the RF models were better than the results of the other three models. The Kappa coefficient larger than 0.4 only existed in RF models using variable combinations 8-12 with overall accuracy from 52% to 61% assessed by field point data; the Kappa coefficient larger than 0.56 only existed in the RF models using variable combinations 8-9 and 11-12 with overall

accuracy from 57% to 61% assessed by field point data; the highest Kappa coefficient of 0.60 with the highest overall accuracy of 61% existed in the RF models using variable combination 8 (Table 5, Fig. 4). When assessed by the VMC, the Kappa coefficient in all models were less than 0.4.

Important variables

For the RF models, 9 important variables in the top 10 most important variables were the same for different vegetation levels, i.e., 4 climate variables (annual mean temperature, max temperature of warmest month, min temperature of coldest month and annual precipitation), 2 geospatial variables (elevation and slope), and 3 summer spectral variables (surface albedo of bands 2 and band 6 and the brightness index). For the DT models, 6 important variables in the top 10 most important variables were the same for different vegetation levels, i.e., 4 climate variables (annual mean temperature, max temperature of warmest month, min temperature of coldest month and annual precipitation), 1 geospatial variable (elevation) and 1 winter spectral variable (winter surface albedo of band 4) (Table 6). More winter spectral variables were included in the DT models, while more summer spectral variables were included in the RF models, indicating that the winter and summer spectral variables were effective for the DT and RF models, respectively (Table 6).

Discussion

Vegetation classification units

Vegetation classification, a complex, multi-level, non-linear system, is one of the complex problems in vegetation ecology research; the higher demanding classification method would not only accurately classify vegetation in low levels of a classification unit but also describe the diversity of ecosystems, especially for the situation of global change (Faber-Langendoen et al., 2014). Because plants in different vegetation classification units have different spectral characteristics, climatic conditions, etc. and because these spectral characteristics and climatic conditions are the basis for the simulation of vegetation distribution, the same models using same variables to simulate the vegetation distribution of different classification units may result in different classification accuracies (Dobrowski et al., 2008), and it can even be said that the map accuracy is a function of the classification system and categories (Muchoney et al., 2000).

There are some reports on vegetation distribution simulation using different vegetation classifications systems. Plant functional types (PFTs), where PFT is defined as a set of plants sharing the same response to a perturbation and having similar effects on the dominant ecosystem processes, are often used to simulate the vegetation distribution, and examples include the Biome and Box system models (Box, 1981; Box, 1996; Dormann & Woodin, 2002; Weng, 2004); further, the simulation results using Biome and Box systems were good (Box, 1981; Song, Zhou & Ouyang, 2005; Weng & Zhou, 2006). The MAPSS model was also used to simulate the vegetation distribution through vegetation life forms, obtained vegetation heat, leaf area index, leaf morphology and leaf longevity (Zhao et al., 2002). The Holdridge life zone

model was used by some researchers to study the potential vegetation distribution, and the simulated potential distribution of vegetation agreed well with the vegetation pattern (Zheng et al., 2006). The IGBP classification system was applied to simulate the vegetation distribution at the regional scale, and the map estimate was upwards of 80% (Muchoney et al., 2000). Unlike previous research, our research used machine learning models and a hierarchical classification system in the VMC, i.e., the highest classification level (vegetation groups) mainly stems from the appearance of communities, the second highest level (vegetation types) mainly stems from the appearance of communities and climate, and the middle classification level (vegetation formations and subformations) stems from dominant species, to show the predictive ability of different models in various classification levels in the region affected by high socioeconomic development. In general, the accuracy of vegetation distribution simulations in higher classification units should be higher. However, in this study, the accuracy of the vegetation distribution simulation in the vegetation type was the highest compared with the other two classification units (Tables 3-5).

Performances of the different models

Our interest in vegetation distribution modeling is driven by our need to forecast and respond to the impact of management actions or environmental changes on vegetation patterns from local to global scales. Making predictions of vegetation distributions can help people to understand the relationship between plants and their abiotic and biotic environments, which is the basis of ecology (Franklin, 2010). To regulate ecosystem service functions and benefit human beings, people can design vegetation distributions according to factors that affect the distribution and abundance of vegetation based on the prediction results of important patterns and trends extracted from vast amounts of data (Hastie, Tibshirani & Friedman, 2009). Therefore, a number of methods related to statistics and machine learning have been used recently with mapped biological and environmental data to model vegetation distributions over large spatial extents at higher resolutions, and vegetation classification has become a widely used method in ecology (Cutler et al., 2007; Franklin, 2010). Even there are many new image classification methods, they are rarely used in the same classification research, especially when combined with environment variables (Li et al., 2014).

In this research, the RF models performed better than the DT, SVM, and MLC models in three classification levels. This finding is consistent with the idea of some researchers who also believe that the RF model performs better when modeling the vegetation distribution compared with other methods (Franklin, 2010). The DT model divides the data into subgroups that are homogeneous according to the ranges of predictor variables values. The DT model is generally able to handle a large number of independent variables, and the time to build a tree model is shorter than that of other methods. However, the DT model is somewhat unstable for vegetation distribution modeling and has a lower classification accuracy. The RF model generates a large number of independent trees through data subsets, and each split in every tree model is developed using a random subset of predictor variables; in general, the effect of the RF model is

better than that of the DT model because the RF model was developed based on the DT model (Franklin, 2010). The SVM model is developed from statistical learning that iteratively locates boundaries of potentially nonlinear or multiple linear between individual training points to discriminate class samples and then optimizes the separation of boundaries between two class samples. The aim of the MLC model is to maximize the overall probability that a pixel is assigned to a class correctly; however, the requirement of the MLC model for a large number of training samples limits its application (Sesnie et al., 2010). Some researchers have shown that the classification accuracies were higher when using the SVM classifier compared to the MLC model (Pal & Mather, 2005; Boyd, Sanchez-Hernandez &Foody, 2006; Sanchez-Hernandez, Boyd & Foody, 2007; Sesnie et al., 2010), and the DT method provided classifications that were significantly more accurate than those of the MLC model in some cases because of the less demands of the DT method (Boyd, Sanchez-Hernandez &Foody, 2006). Some other researchers have shown that the accuracies simulated by the RF and SVM models were actually very close, with accuracies of 65.3% and 66.6%, respectively (Sesnie et al., 2010), and some researchers have found that RF, MLC, DT and SVM models performed similarly and reasonably well when they simulated land use classification (Li et al., 2014). In addition to the above methods, an artificial neural network was implemented at the regional scale, with classification accuracies of 60%-80% (Muchoney et al., 2000; Haslem et al., 2010); in the Arctic, this method provided the most accurate mapping of vegetation types (Langford et al., 2019). The reasons for the similar and good results of these models may be as follows: the differences between their classification objects were relatively large; they used sufficiently representative training samples; and their input variables were appropriate. In this research, the SVM and MLC models only output the simulated results of the variable combinations 1-6. The reason might be the poor separability of the training samples; the models could not recognize the training points and their vegetation categories (Jarnevich et al., 2015). There are many types of vegetation in the Jing-Jin-Ji region, and the distribution area of some types is very small, so the training points selected from these types may not be satisfactory. Sufficient training points for these types may be needed in future research through field surveys. To determine a more suitable model, more models, such as one that is suitable for modelling the global vegetation distribution in the future, should be developed and tested (Jiang et al., 2012).

Important variables

Variable selection is directly related to the capacity of a vegetation distribution model to capture important vegetation environmental requirements, and these variables include temperature, precipitation, and topography (Mod et al., 2016). In addition to the environmental variables, some spectral variables are used as input variables, but the use of too many spectral variables can actually decrease the discrimination accuracy, and some spectral variables that reflect vegetation information should be selected, such as variables related to the visible spectrum, infrared spectrum, and vegetation indices (Price, Guo & Stiles, 2002, Zhou et al., 2016). Different variables respond to different information. Spectral variables reflect directly the information of

land surface objects, while geospatial and climatic variables indicate information about the vegetation environment.

Elevation, an important variable for vegetation distribution models, especially in regions with large elevation differences, has long been used to enhance map accuracies (Franklin, 1995; Dobrowski et al. 2008; Oke & Thompson, 2015; Zhou et al., 2016). Sesnie et al. (2010) demonstrated that adding elevation as an additional predictor variable dramatically improved the accuracies of SVM and RF models to levels >80% for most forest types. At the same elevation, slopes with different aspects have very different temperatures of soil and vegetation (Gunton, Polce & Kunin, 2015; Mod et al., 2016). Dobrowski et al. (2008) highlighted the importance of slope and aspect when mapping vegetation communities in the Sierra Nevada. Elevation and slope were also important variables in this research (Table 6). Different types of vegetation have different requirements for annual precipitation and temperature, and they also have different tolerances to extreme heat and cold. The importance of these climate variables (annual mean temperature, extreme temperature and annual precipitation) has been validated in other studies (Sesnie et al., 2008; Zhou et al., 2016) and was tested in our research. Three surface albedo indices (the summer surface albedo of bands 2 and band 6 and the winter surface albedo of band 4) were important variables in this study. These indices were the near-infrared, short-wave infrared and green bands in the Modis images. Radiation in the near infrared (760–900 nm) and green wavelengths (520-600 nm) is strongly reflected by leaf cellular structures and chlorophyll, respectively, and the absorption rate of short-wave infrared radiation (1638-1652 nm) is greatly increased due to the influence of the water content of green vegetation (Peng et al., 2002). Near infrared radiation is always selected in discriminate analysis as the best discriminating variable (Price, Guo & Stiles, 2002). Sesnie et al. (2010) combined elevation and spectral band data to increase the classification accuracy to a satisfactory level for most forest types. De Colstoun et al. (2003) got high accuracies (80%) for classifying coniferous, temperate broad-leaf, and mixed forest types with Landsat ETM+ bands. Other important vegetation index variables have been used in other studies (Price, Guo & Stiles, 2002; Zhou et al., 2016), depending on the specific study area and data.

The input variables used in a vegetation distribution model should not be limited to the input variables in this study. Some ecophysiologically meaningful predictors might be considered, such as soil moisture, soil pH, and soil nutrients. In addition, some other factors, such as actual light, disturbance, biotic interactions, land use, and bioclimatic information, might be incorporated into vegetation distribution models (Dobrowski et al., 2008; Sesnie et al., 2010; Mod et al., 2016). We suggest that more effort should be made to build more ecophysiologically sound vegetation distribution models, which would require a collaborative effort among the ecological, geographical and environmental science (Mod et al., 2016).

Other factors affecting the accuracy of classification

In addition to the classification units, models, and input variables, other factors influence the accuracy of classification, such as algorithm error, image data and other objective reasons (Li et

al., 2014). We must acknowledge the existence of errors in random sample selection and algorithms including models and data preprocessing. The sources of remote sensing data are different, and the date and processing of selected images are different, which results in different values of remote sensing images and even different accuracies of the simulated results (Price, Guo & Stiles, 2002). Furthermore, remote sensing images with high spectral and spatial resolutions provide rich spectral and ground information. They can improve the predictive ability of the vegetation distribution model to some extent (Peng et al., 2002). However, the use of high spectral and spatial resolution images is creating a greater demand for access to these data, larger computer storage capacities and faster data processors (Price, Guo & Stiles, 2002), which is why high spectral and spatial resolution images were not used in this study. Moreover, some vegetation types such as cultivated vegetation and shelter forests in the Jing-Jin-Ji region are greatly affected by humans. Their water-heat conditions are artificially controlled, which is inconsistent with the climate variable inputs in this study and may reduce the predictive ability of the vegetation distribution model. Finally, the VMC used for model assessment in this study was published in 2007, and no new study has been published in the last 10 years. The actual vegetation may not be completely consistent with the VMC in the Jing-Jin-Ji region, which may cause the accuracy of the assessment by the VMC to be relatively low.

Conclusions

RF models could well simulate the vegetation distribution in all three levels, i.e., the vegetation group, vegetation type, and formation and subformation. The DT, SVM and MLC models could not simulate the distribution in all three levels. Based on the Kappa coefficient, the RF model was the best model for simulating the vegetation distribution in the Jing-Jin-Ji region. Five variables, including 4 climate variables (annual mean temperature, max temperature of warmest month, min temperature of coldest month and annual precipitation) and 1 geospatial variable (elevation), were the most important to increase the accuracy of vegetation classification, and the winter surface albedo of band 4, the slope, and the three summer spectral variables (the summer surface albedo of bands 2 and 6 and the brightness index) could increase the accuracy of vegetation classification to some extent.

Acknowledgements

We thank two anonymous reviewers and the editor for their effort to review this manuscript.

References

Bannari A, Morin D, Bonn F. 1995. A review of vegetation indices. Remote Sensing Reviews 13(1-2):95-120 DOI:10.1080/02757259509532298.

Bie SW, Beckett PHT. 1973. Comparison of four independent soil surveys by air-photo interpretation, paphos area (cyprus). Photogrammetria 29(6):189-202 DOI:10.1016/0031-8663(73)90001-x.

Burai P, Deak B, Valko O, Tomor T. 2015. Classification of Herbaceous Vegetation Using Airborne Hyperspectral Imagery. Remote Sensing 7(2):2046-2066 DOI:10.3390/rs70202046.

Box EO. 1981. Macroclimate and plant forms: An introduction to predictive modeling in phytogeography (Tasks for vegetation science 1). London: DR. W. Junk Publishers.

Box EO. 1996. Plant functional types and climate at the global scale. Journal of vegetation science 7(3):309-320 DOI:10.2307/3236274.

Boyd DS, Sanchez-Hernandez C, Foody GM. 2006. Mapping a specific class for priority habitats monitoring from satellite sensor data. International Journal of Remote Sensing 27(13):2631-2644 DOI:10.1080/01431160600554348.

Chang KT, Chen JF, Lian L. 2014. Introduction to Geographic Information Systems (Seventh Edition). Beijing: Publishing house of electronics industry.

Chen LY, Li H, Zhang PJ, Zhao X, Zhou LH, Liu TY, Hu HF, Bai YF, Shen HH, Fang JY. 2015. Climate and native grassland vegetation as drivers of the community structures of shrub-encroached grasslands in Inner Mongolia, China. Landscape Ecology 30(9):1627-1641 DOI:10.1007/s10980-014-0044-9.

Cohen WB, Goward SN. 2004. Landsat's role in ecological applications of remote sensing. Bioscience 54(6):535-545 DOI:10.1641/0006-3568(2004)054[0535:lrieao]2.0.co;2.

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ. 2007. Random Forests for Classification in Ecology. Ecology 88(11):2783-2792 DOI:10.1890/07-0539.1.

Deng SB. 2010. ENVI remote sensing image processing method. Beijing: Science press.

de Colstoun ECB, Story MH, Thompson C, Commisso K, Smith TG, Irons JR. 2003. National Park vegetation mapping using multitemporal Landsat 7 data and a decision tree classifier. Remote Sensing of Environment 85(3):316-327 DOI:10.1016/s0034-4257(03)00010-5.

Dilts TE, Weisberg PJ, Dencker CM, Chambers JC. 2015. Functionally relevant climate variables for arid lands: a climatic water deficit approach for modelling desert shrub distributions. Journal of Biogeography 42(10):1986-1997 DOI:10.1111/jbi.12561.

Dobrowski SZ, Safford HD, Cheng YB, Ustin SL. 2008. Mapping mountain vegetation using species distribution modeling, image-based texture analysis, and object-based classification. Applied Vegetation Science 11(4):499-508 DOI:10.3170/2008-7-18560.

Dormann CF, Woodin SJ. 2002. Climate change in the Arctic: using plant functional types in a meta-analysis of field experiments. Functional Ecology 16(1):4-17 DOI:10.1046/j.0269-8463.2001.00596.x.

Editorial Committee of Vegetation Map of China, the Chinese Academy of Sciences. 2007. The Vegetation Map of the People’s Republic of China (1:1 000 000). Beijing: Geological Publishing House.

Editorial Committee of Vegetation Map of China, the Chinese Academy of Sciences. 2007b. The Chinese vegetation and its geographical pattern. Beijing: Geological Publishing House.

Faber-Langendoen D, Keeler-Wolf T, Meidinger D, Tart D, Hoagland B, Josse C, Navarro G, Ponomarenko S, Saucier J-P, Weakley A, Comer P. 2014. EcoVeg: a new approach to

vegetation description and classification. Ecological Monographs 84(4):533-561 DOI:10.1890/13-2334.1.
Fick SE, Hijmans RJ. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology 37(12):4302-4315 DOI:10.1002/joc.5086.
Franklin J. 1995. Predictive vegetation mapping: Geographic modelling of biospatial patterns in relation to environmental gradients. Progress in Physical Geography 19(4):474-499 DOI:10.1177/030913339501900403.
Franklin J. 2010. Mapping species distributions: spatial inference and prediction. Cambridge: Cambridge University Press.
Gislason PO, Benediktsson JA, Sveinsson JR. 2006. Random Forests for land cover classification. Pattern Recognition Letters 27(4):294-300 DOI:10.1016/j.patrec.2005.08.011.
Gunton RM, Polce C, Kunin WE. 2015. Predicting ground temperatures across European landscapes. Methods in Ecology and Evolution 6(5):532-542 DOI:10.1111/2041-210x.12355.
Guisan A, Zimmermann NE. 2000. Predictive habitat distribution models in ecology. Ecological modelling 135:147-186 DOI:10.1016/s0304-3800(00)00354-9.
Haslem A, Callister KE, Avitabile SC, Griffioen PA, Kelly LT, Nimmo DG, Spence-Bailey LM, Taylor RS, Watson SJ, Brown L, Bennett AF, Clarke MF. 2010. A framework for mapping vegetation over broad spatial extents: A technique to aid land management across jurisdictional boundaries. Landscape and Urban Planning 97(4):296-305 DOI:10.1016/j.landurbplan.2010.07.002.
Hastie T, Tibshirani R, Friedman J. 2009. The elements of statistical learning: Data Mining, Inference, and Prediction (Second Edition). Berlin: Springer.
Hansen MC, Potapov PV, Moore R, Hancher M, Turubanova SA, Tyukavina A, Thau D, Stehman SV, Goetz SJ, Loveland TR, Kommareddy A, Egorov A, Chini L, Justice CO, Townshend JRG. 2013. High-Resolution Global Maps of 21st-Century Forest Cover Change. Science 342(6160):850-853 DOI:10.1126/science.1244693.
Jarnevich CS, Stohlgren TJ, Kumar S, Morisette JT, Holcombe TR. 2015. Caveats for correlative species distribution modeling. Ecological Informatics 29(6-15) DOI:10.1016/j.ecoinf.2015.06.007.
Jiang, H., Zhao, D., Cai, Y., et al., 2012. A Method for Application of Classification Tree Models to Map Aquatic Vegetation Using Remotely Sensed Images from Different Sensors and Dates. Sensors. 12(9): 12437-12454.
Jiang H, Zhao D, Cai Y, An S. 2012. A Method for Application of Classification Tree Models to Map Aquatic Vegetation Using Remotely Sensed Images from Different Sensors and Dates. Sensors 12(9):12437-12454 DOI:10.3390/s120912437.
Kabacoff RI. 2016. R in action: data analysis and graphics with R (Second Edition). Beijing: Posts & Telecom Press.

557 Landis JR, Koch GG. 1977. The Measurement of observer agreement for categorical data.
558 Biometrics 33(1):159-174 DOI:10.2307/2529310.
559 Langford ZL, Kumar J, Hoffman FM, Breen AL, Iversen CM. 2019. Arctic Vegetation Mapping
560 Using Unsupervised Training Datasets and Convolutional Neural Networks. Remote
561 Sensing 11(1):69 DOI:10.3390/rs11010069.
562 Lany NK, Zarnetske PL, Finley AO, McCullough DG. 2019. Complementary strengths of
563 spatially-explicit and multi-species distribution models. Ecography 42:1-11
564 DOI:10.1111/ecog.04728.
565 Li C, Wang J, Wang L, Hu L, Gong P. 2014. Comparison of Classification Algorithms and
566 Training Sample Sizes in Urban Land Classification with Landsat Thematic Mapper
567 Imagery. Remote Sensing 6(2):964-983 DOI:10.3390/rs6020964.
568 Muchoney D, Borak J, Chi H, Friedl M, Gopal S, Hodges J, Morrow N, Strahler A. 2010.
569 Application of the MODIS global supervised classification model to vegetation and land
570 cover mapping of Central America. International Journal of Remote Sensing 21(6-7):1115-
571 1138 DOI:10.1080/014311600210100.
572 Mod HK, Scherrer D, Luoto M, Guisan A. 2016. What we use is not what we know:
573 environmental predictors in plant distribution models. Journal of Vegetation Science
574 27(6):1308-1322 DOI:10.1111/jvs.12444.
575 Newell CL, Leathwick JR. 2005. Mapping Hurunui forest community distribution, using
576 computer models (Science for Conservation. 251). Wellington: Department of
577 Conservation.
578 Oke OA, Thompson KA. 2015. Distribution models for mountain plant species: The value of
579 elevation. Ecological Modelling 301:72-77 DOI:10.1016/j.ecolmodel.2015.01.019.
580 Pal M, Mather PM. 2005. Support vector machines for classification in remote sensing.
581 International Journal of Remote Sensing 26(5):1007-1011
582 DOI:10.1080/01431160512331314083.
583 Peng WL, Bai ZP, Liu XN, Cao T. 2002. Introduction to remote sensing. Beijing: Higher
584 education press.
585 Pfeffer K, Pebesma EJ, Burrough PA. 2003. Mapping alpine vegetation using vegetation
586 observations and topographic attributes. Landscape Ecology 18(8):759-776
587 DOI:10.1023/B:LAND.0000014471.78787.d0.
588 Price KP, Guo XL, Stiles JM. 2002. Optimal Landsat TM band combinations and vegetation
589 indices for discrimination of six grassland types in eastern Kansas. International Journal of
590 Remote Sensing 23(23):5031-5042 DOI:10.1080/01431160210121764.
591 Sanchez-Hernandez C, Boyd DS, Foody GM. 2007. Mapping specific habitats from remotely
592 sensed imagery: Support vector machine and support vector data description based
593 classification of coastal saltmarsh habitats. Ecological Informatics 2(2):83-88
594 DOI:10.1016/j.ecoinf.2007.04.003.
595 Sesnie SE, Finegan B, Gessler PE, Thessler S, Bendana ZR, Smith AMS. 2010. The
596 multispectral separability of Costa Rican rainforest types with support vector machines and

Random Forest decision trees. International Journal of Remote Sensing 31(11):2885-2909 DOI:10.1080/01431160903140803.

Sesnie SE, Gessler PE, Finegan B, Thessler S. 2008. Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments. Remote Sensing of Environment 112(5):2145-2159 DOI:10.1016/j.rse.2007.08.025.

Sitch S, Smith B, Prentice IC, Arneth A, Bondeau A, Cramer W, Kaplan JO, Levis S, Lucht W, Sykes MT, Thonicke K, Venevsky S. 2003. Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. Global Change Biology 9(2):161-185 DOI:10.1046/j.1365-2486.2003.00569.x.

Song MH, Zhou CP, Ouyang H. 2005. Simulated distribution of vegetation types in response to climate change on the Tibetan Plateau. Journal of Vegetation Science 16(3):341-350 DOI:10.1111/j.1654-1103.2005.tb02372.x.

van der Linden S, Rabe A, Held M, Jakimow B, Leitao PJ, Okujeni A, Schwieder M, Suess S, Hostert P. 2015. The EnMAP-Box-A Toolbox and Application Programming Interface for EnMAP Data Processing. Remote Sensing 7(9):11249-11266 DOI:10.3390/rs70911249.

Wang L, Gong BL. 2018. Collaborative Governance of Ecological Space in Beijing-Tianjin-Hebei Region. Journal of Tianjin administration institute 20(5):38-44 DOI:10.16326/j.cnki.1008-7168.2018.05.005. (Chinese)

Wang X, Liu G, Coscieme L, Giannetti BF, Hao Y, Zhang Y, Brown MT. 2019. Study on the emergy-based thermodynamic geography of the Jing-Jin-Ji region: Combined multivariate statistical data with DMSP-OLS nighttime lights data. Ecological Modelling 397:1-15 DOI:10.1016/j.ecolmodel.2019.01.021.

Wehkamp J, Pietsch SA, Fuss S, Gusti M, Reuter WH, Koch N, Kindermann G, Kraxner F. 2018. Accounting for institutional quality in global forest modeling. Environmental Modelling & Software 102:250-259 DOI:10.1016/j.envsoft.2018.01.020.

Weng ES, Zhou GS. 2006. Modeling distribution changes of vegetation in China under future climate change. Environmental Modeling & Assessment 11(1):45-58 DOI:10.1007/s10666-005-9019-1.

Xie YC, Sha ZY, Yu M. 2008. Remote sensing imagery in vegetation mapping: a review. Journal of Plant Ecology 1(1):9-23 DOI:10.1093/jpe/rtm005.

Weng, ES., 2004. Division of plant functional type-biome in China. D. Thesis, Institute of botany, the Chinese academy of science.

Zhang G, Biradar CM, Xiao X, Dong J, Zhou Y, Qin Y, Zhang Y, Liu F, Ding M, Thomas RJ. 2018. Exacerbated grassland degradation and desertification in Central Asia during 2000-2014. Ecological Applications 28(2):442-456 DOI:10.1002/eap.1660.

Zhang WT, Dong W. 2013. SPSS statistical analysis advanced tutorial (Second Edition). Beijing: Higher education press.

Zhang WT, Dong W. 2017. SPSS statistical analysis advanced tutorial (Third Edition). Beijing: Higher education press.

637 Zhang Z, De Clercq E, Ou XK, De Wulf R, Verbeke L. 2008. Mapping dominant vegetation
638 communities at Meili Snow Mountain, Yunnan Province, China using satellite imagery and
639 plant community data. Geocarto International 23(2):135-153
640 DOI:10.1080/10106040701337410.
641 Zhao MS, Neilson RP, Yan XD, Dong WJ. 2002. Modelling the vegetation of China under
642 changing climate. Acta geographica sinica 57(1):28-38. DOI:CNKI:SUN:DLXB.0.2002-01-
643 003. (Chinese)
644 Zhao X, Su Y, Hu T, Chen L, Gao S, Wang R, Jin S, Guo Q. 2018. A global corrected SRTM
645 DEM product for vegetated areas. Remote Sensing Letters 9(4):393-402
646 DOI:10.1080/2150704x.2018.1425560.
647 Zheng Y, Xie Z, Jiang L, Shimizu H, Drake S. 2006. Changes in Holdridge Life Zone diversity
648 in the Xinjiang Uygur Autonomous Region (XUAR) of China over the past 40 years.
649 Journal of Arid Environments 66(1):113-126. DOI:10.1016/j.jaridenv.2005.09.005.
650 Zhou J, Lai L, Guan T, Cai W, Gao N, Zhang X, Yang D, Cong Z, Zheng Y. 2016. Comparison
651 modeling for alpine vegetation distribution in an arid area. Environmental Monitoring and
652 Assessment 188(7):408 DOI:10.1007/s10661-016-5417-x.
653

Table 1(on next page)

Classification units of the vegetation of China.

Table 1:

Classification units of the vegetation of China

Vegetation groups	Vegetation types	Formations and sub-formations
0. No vegetation	0 No vegetation	0 No vegetation
1. Needleleaf forest	1 Temperate needleleaf forest	1 *Pinus tabulaeformis* forest
2. Broadleaf forest	2 Temperate broadleaf deciduous forest	2 *Quercus mongolica* forest
		3 *Quercus liaotungensis* forest
		4 *Quercus variabilis* forest
		5 *Robinia pseudoacacia* forest
		6 *Salix matsudana* forest
		7 *Populus davidiana* forest
		8 *Betula platyphylla* forest
3. Scrub	3 Temperate broadleaf deciduous scrub	9 *Corylus heterophylla* scrub
		10 *Lespedeza bicolor* scrub
		11 *Prunus armeniaca var. ansa* scrub
		12 *Vitex negundo* var. *heterophylla, Zizyphus jujuba* var. *spinosa* scrub
		13 *Cotinus coggygria* var. *cinerea* scrub
		14 *Spiraea spp.* scrub
		15 *Ostryopsis davidiana* scrub
4. Steppe	4 Temperate grass-forb meadow steppe	16 *Stipa baicalensis,* forb meadow steppe
		17 *Filifolium sibiricum,* grass-forb meadow steppe
	5 Temperate needlegrass arid steppe	18 *Aneurolepidium chinense*, needlegrass steppe
		19 *Stipa krylovii* steppe
		20 *Stipa bungiana* steppe
		21 *Thymus mongolicus,* needlegrass steppe
5. Grass-forb community	6 Temperate grass-forb community	22 *Bothriochloa ischaemum* community
		23 *Bothriochloa ischaemum* community
		24 *Vitex negundo* var. *heterophylla, Zizyphus jujuba var. spinosus*, *Bothriochloa ischaemum* scrub and grass community
		25 *Vitex negundo* var. *heterophylla, Zizyphus jujuba* var. *spinosus, Themeda triandra* var. *japonica* scrub and grass community
6. Meadow	7 Temperate grass and forb meadow	26 *Arundinella hirta, Spodiopogon sibiricus*, forb meadow
		27 *Carex spp.,* forb meadow
	8 Temperate grass and forb holophytic meadow	28 *Achnatherum splendens* holophytic meadow
		29 *Suaeda glauca* holophytic meadow
7. Swamp	9 Cold-temperate	30 *Phragmites communis* swamp

	and temperate swamp	
8. Cultural vegetation	10 One crop annually and cold-resistant economic crops	31 Spring wheat, naked oats, buckwheat, potatoes; flux
	11 One crop annually, cold-resistant economic crops and deciduous orchards	32 Coarse grains
	12 Three crops two years and two crops annually non irrigation, deciduous orchards	33 Winter wheat, coarse grains
		34 Coarse grains
		35 Rice
		36 Winter wheat, corn, cotton
		37 Apple, pear orchard
		38 Winter wheat, corn, Chinese sorghum, sweet potatoes; cotton, tabacco, peanut, sesame; apple, pear, hauthorn, persimmon, walnut, pomegranat, grape
		39 Winter wheat, coarse grains (loamy soil)

Table 2(on next page)

The vegetation indices.

1 **Table 2:**

2 **The vegetation indices**

Indices	Abbreviation	Formula
Ratio vegetation index	RVI	NIR/Red
Brightness index	BI	0.2909Blue + 0.2493Green + 0.4806Red + 0.5568NIR + 0.4438SWIR1 + 0.1706SWIR2
Green vegetation index	GI	-0.2728Blue - 0.2174Green-0.5508Red + 0.7221NIR + 0.0733SWIR1 - 0.1648SWIR2
Wetness index	WI	0.1446Blue + 0.1761Green + 0.3322Red + 0.3396NIR - 0.6210SWIR1 - 0.4186SWIR2
Differenced vegetation index	DVI	NIR - Red
Green ratio	GR	NIR/Green
MIR ratio	MR	NIR/SWIR1
Soil-adjusted vegetation index	SAVI	(1.5(NIR - Red))/((NIR + Red + 0.5))
Optimization of soil-adjusted vegetation index	OSAVI	(1.16(NIR - Red))/((NIR + Red + 0.16))
Atmospherically resistant vegetation index	ARVI	(NIR - (2*Red - Blue))/(NIR + (2*Red - Blue))
Normalized difference vegetation index	NDVI	(NIR - Red)/(NIR + Red)
Enhanced vegetation index	EVI	2.5[(NIR - Red)/(NIR + 6*Red - 7.5Blue + 1)]
Normalized difference tillage index	NDTI	(SWIR1-SWIR2)/(SWIR1 + SWIR2)
Normalized difference senescent vegetation index	NDSVI	(SWIR1-Red)/(SWIR1 + Red)

3

4

Table 3(on next page)

Model assessment of vegetation groups by field point data and VMC.

Variable combinations: 1, 1-7 land surface albedos for summer; 2, 1-7 land surface albedos for winter; 3, 1-7 land surface albedos for summer and winter; 4, vegetation indices for summer; 5, vegetation indices for winter; 6, vegetation indices for summer and winter; 7, geospatial variables; 8, climate variables; 9, geospatial variables and Climate variables; 10, all 49 variables; 11, DT10, the top 10 important variables in decision tree using all variables; 12, RF10, the top 10 important variables in random forest using all variables. VMC, the Vegetation Map of the People's Republic of China. **, the kappa coefficient lager than 0.56; *, the kappa coefficient larger than 0.4 and less than 0.56. OA, Overall accuracy; KC, Kappa coefficient.

Table 3:

Model assessment of vegetation groups by field point data and VMC.

Variable combinations: 1, 1-7 land surface albedos for summer; 2, 1-7 land surface albedos for winter; 3, 1-7 land surface albedos for summer and winter; 4, vegetation indices for summer; 5, vegetation indices for winter; 6, vegetation indices for summer and winter; 7, geospatial variables; 8, climate variables; 9, geospatial variables and Climate variables; 10, all 49 variables; 11, DT10, the top 10 important variables in decision tree using all variables; 12, RF10, the top 10 important variables in random forest using all variables. VMC, the Vegetation Map of the People's Republic of China. **, the kappa coefficient lager than 0.56; *, the kappa coefficient larger than 0.4 and less than 0.56. OA, Overall accuracy; KC, Kappa coefficient.

Variable combinations	Decision tree				Random forest				Support vector machine				Maximum likelihood classification			
	Point data		VMC		Point data		VMC		Point data		VMC		Point data		VMC	
	OA	KC	OA	KC	OA	KC	OA	KC	OA	KC	OA	KC	OA	KC	OA	KC
1	37%	0.22	51%	0.19	48%	0.37	46%	0.21	41%	0.27	58%	0.30	33%	0.19	39%	0.17
2	31%	0.16	47%	0.20	43%	0.31	48%	0.23	39%	0.25	57%	0.30	21%	0.12	16%	0.07
3	39%	0.26	50%	0.23	51%	0.41*	52%	0.28	44%	0.31	57%	0.32	41%	0.30	44%	0.24
4	34%	0.18	54%	0.21	45%	0.33	44%	0.19	41%	0.28	55%	0.25	11%	0.06	6%	0.03
5	32%	0.16	54%	0.23	44%	0.33	48%	0.24	39%	0.26	58%	0.30	12%	0.08	5%	0.04
6	38%	0.25	50%	0.22	51%	0.41*	54%	0.30	47%	0.35	60%	0.34	3%	0.02	1%	0.01
7	42%	0.31	45%	0.26	50%	0.40*	50%	0.28								
8	21%	0.00	67%	0.00	71%	0.65**	55%	0.35								
9	24%	0.13	47%	0.21	71%	0.65**	56%	0.37								
10	47%	0.37	34%	0.20	67%	0.60**	57%	0.38								
11	46%	0.36	36%	0.21	69%	0.63**	55%	0.36								
12	46%	0.36	36%	0.21	72%	0.67**	56%	0.37	□	□	□	□	□	□	□	□

Table 4(on next page)

Model assessment of vegetation types by field point data and VMC.

The Abbreviations were same with Table 3.

Table 4:

Model assessment of vegetation types by field point data and VMC.

The Abbreviations were same with Table 3.

Variable combinations	Decision tree				Random forest				Support vector machine				Maximum likelihood classification			
	Point data		VMC		Point data		VMC		Point data		VMC		Point data		VMC	
	OA	KC	OA	KC	OA	KC	OA	KC	OA	KC	OA	KC	OA	KC	OA	KC
1	32%	0.18	40%	0.18	43%	0.34	38%	0.21	38%	0.26	46%	0.27	12%	0.07	10%	0.03
2	22%	0.06	12%	0.03	41%	0.30	41%	0.24	35%	0.22	47%	0.29	11%	0.06	22%	0.07
3	36%	0.24	43%	0.24	47%	0.37	44%	0.28	41%	0.31	42%	0.27	23%	0.17	30%	0.16
4	31%	0.16	42%	0.17	40%	0.30	33%	0.17	37%	0.24	42%	0.21	19%	0.12	11%	0.05
5	28%	0.14	43%	0.23	43%	0.32	41%	0.24	39%	0.27	49%	0.31	7%	0.04	6%	0.03
6	36%	0.24	43%	0.24	47%	0.38	46%	0.30	45%	0.35	48%	0.31	7%	0.05	3%	0.02
7	41%	0.31	49%	0.35	47%	0.38	47%	0.33								
8	39%	0.29	32%	0.18	70%	0.65**	53%	0.40*								
9	37%	0.28	20%	0.12	71%	0.67**	55%	0.43*								
10	39%	0.30	19%	0.12	65%	0.60**	55%	0.43*								
11	44%	0.36	33%	0.22	68%	0.63**	54%	0.42*								
12	45%	0.37	37%	0.25	71%	0.65**	55%	0.41*	□	□	□	□	□	□	□	□

Table 5(on next page)

Model assessment of formations and sub-formations by field point data and VMC.

The Abbreviations were same with Table 3.

Table 5:

Model assessment of formations and sub-formations by field point data and VMC.

The Abbreviations were same with Table 3.

Variable combinations	Decision tree				Random forest				Support vector machine				Maximum likelihood classification			
	Point data		VMC		Point data		VMC		Point data		VMC		Point data		VMC	
	OA	KC	OA	KC	OA	KC	OA	KC	OA	KC	OA	KC	OA	KC	OA	KC
1	20%	0.14	38%	0.12	25%	0.23	9%	0.06	16%	0.14	7%	0.05	10%	0.08	10%	0.06
2	10%	0.08	5%	0.03	26%	0.24	9%	0.06	16%	0.13	8%	0.06	13%	0.10	16%	0.10
3	11%	0.09	5%	0.04	32%	0.30	12%	0.09	25%	0.23	10%	0.08	19%	0.17	16%	0.11
4	10%	0.07	5%	0.03	23%	0.21	8%	0.05	16%	0.13	7%	0.04	5%	0.04	5%	0.03
5	9%	0.07	4%	0.03	26%	0.24	9%	0.07	20%	0.18	9%	0.06	12%	0.10	13%	0.08
6	13%	0.11	4%	0.03	33%	0.31	12%	0.09	27%	0.24	11%	0.09	2%	0.01	0%	0.00
7	11%	0.08	9%	0.07	25%	0.23	12%	0.09								
8	15%	0.13	10%	0.08	61%	0.60**	24%	0.21								
9	39%	0.36	25%	0.21	59%	0.58**	25%	0.22								
10	34%	0.31	28%	0.24	52%	0.51*	22%	0.19								
11	39%	0.36	28%	0.24	57%	0.56**	24%	0.22								
12	40%	0.36	30%	0.27	59%	0.58**	25%	0.22	□	□	□	□	□	□	□	□

Table 6(on next page)

Top ten most important variables of models in the different vegetation classification units.

The abbreviations of indices were shown in Table 2.

Table 6:

Top ten most important variables of models in the different vegetation classification units.

The abbreviations of indices were shown in Table 2.

	Vegetation groups				Vegetation types				Formations and sub-formations			
	Decision tree		Random forest		Decision tree		Random forest		Decision tree		Random forest	
	Important variables	Standardized Importance	Important variables	Normalized importance	Important variables	Standardized Importance	Important variables	Normalized importance	Important variables	Standardized Importance	Important variables	Normalized importance
1	Elevation	1.00	Annual precipitation	1.91	Elevation	1.00	Annual precipitation	1.85	Min temperature of coldest month	1.00	Annual precipitation	1.94
2	Max temperature of warmest month	0.98	Annual mean temperature	1.66	Max temperature of warmest month	0.96	Annual mean temperature	1.65	Annual mean temperature	0.87	Annual mean temperature	1.71
3	Annual mean temperature	0.96	Max temperature of warmest month	1.60	Annual mean temperature	0.91	Max temperature of warmest month	1.63	Elevation	0.85	Max temperature of warmest month	1.61
4	Min temperature of coldest month	0.61	Elevation	1.39	Annual precipitation	0.59	Elevation	1.36	Annual precipitation	0.82	Elevation	1.39
5	Annual precipitation	0.59	Slope	1.36	Win temperature of coldest month	0.58	Slope	1.35	Max temperature of warmest month	0.76	Slope	1.35
6	Winter index WI	0.46	Min temperature of coldest month	1.31	Winter index DVI	0.40	Min temperature of coldest month	1.31	Winter index DVI	0.58	Min temperature of coldest month	1.34
7	Winter index BI	0.46	Summer surface albedo of band 6	1.20	Winter index WI	0.39	Summer surface albedo of band 6	1.18	Winter index GI	0.53	Summer surface albedo of band 6	1.18

8	Winter surface albedo of band 3	0.45	Summer surface albedo of band 2	0.92	Winter index GI	0.38	Summer surface albedo of band 2	0.90	Winter surface albedo of band 2	0.51	Summer indices of BI	0.85
9	Winter surface albedo of band 2	0.45	Summer index BI	0.87	Winter surface albedo of band 4	0.37	Summer indices BI	0.80	summer index GI	0.51	winter surface albedo of band 5	0.84
10	Winter surface albedo of band 4	0.45	Summer index GR	0.84	Winter surface albedo of band 1	0.36	Summer index EVI	0.78	Winter surface albedo of band 4	0.51	Summer surface albedo of band 2	0.84

4

5

Figure 1

The location and DEM of the Jng-Jin-Ji region.

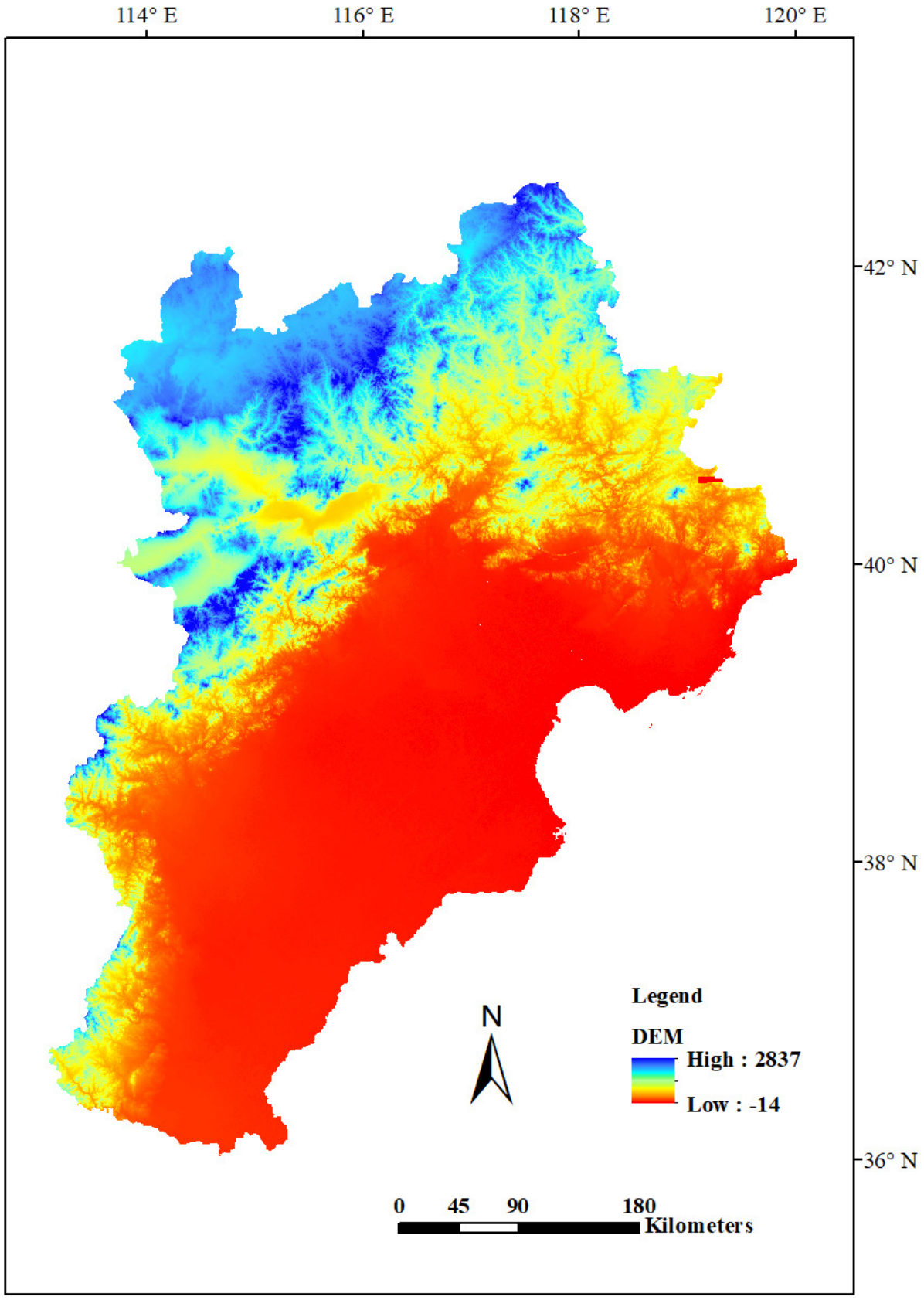
114° E
116° E
118° E
120° E
42° N
40° N
38° N
36° N
N
Legend
DEM
High : 2837
Low : -14
0
45
90
180
Kilometers

Figure 2

The modeling vegetation map of vegetation groups with highest accuracy by Decision tree model (a), Random forest model (b), Support vector machine (c), Maximum likelihood classification (d) and the Vegetation Map of the People's Republic of China in the J

The legend represents vegetation groups shown in Table 1.

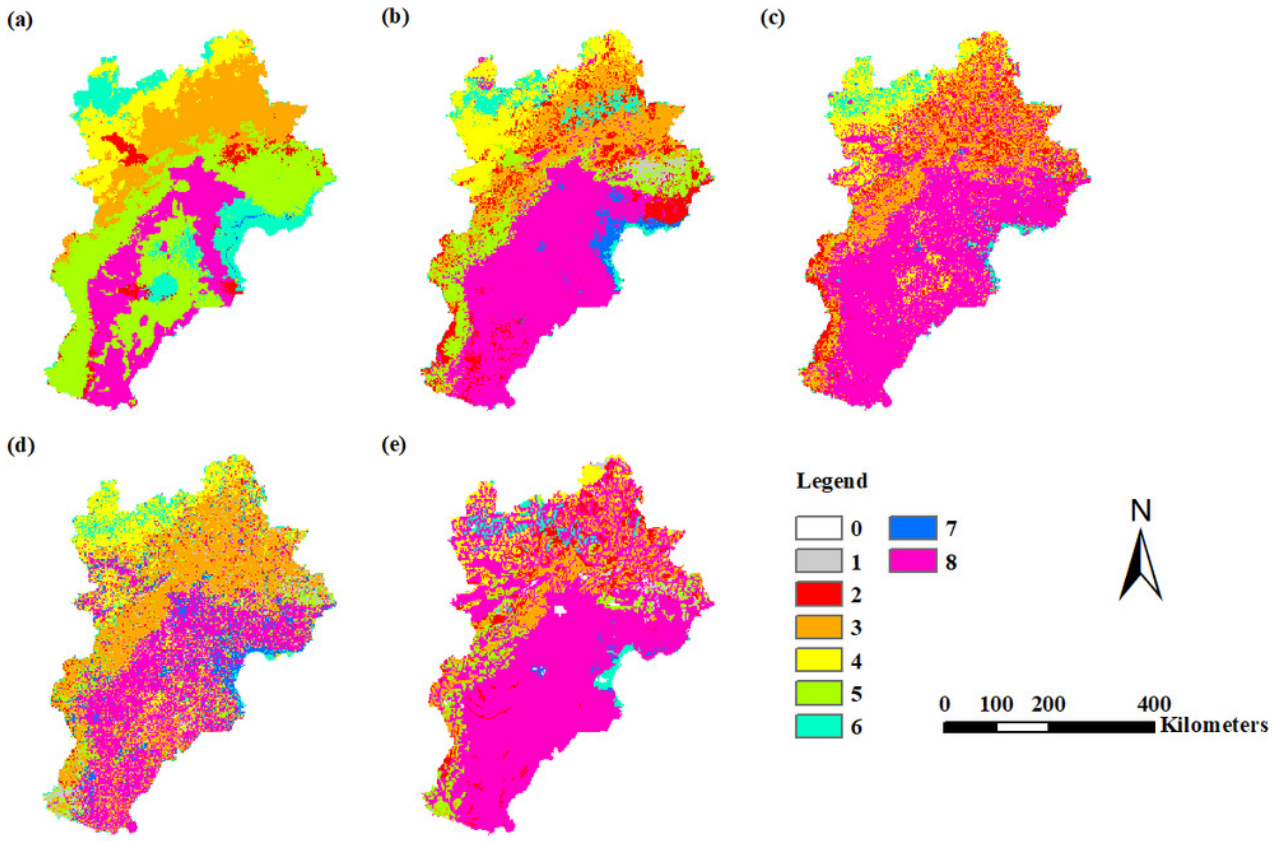

Figure 3

The modeling vegetation map of vegetation types with highest accuracy by Decision tree model (a), Random forest model (b), Support vector machine (c), Maximum likelihood classification (d) and the Vegetation Map of the People's Republic of China in the Ji

The legend represents vegetation types shown in Table 1.

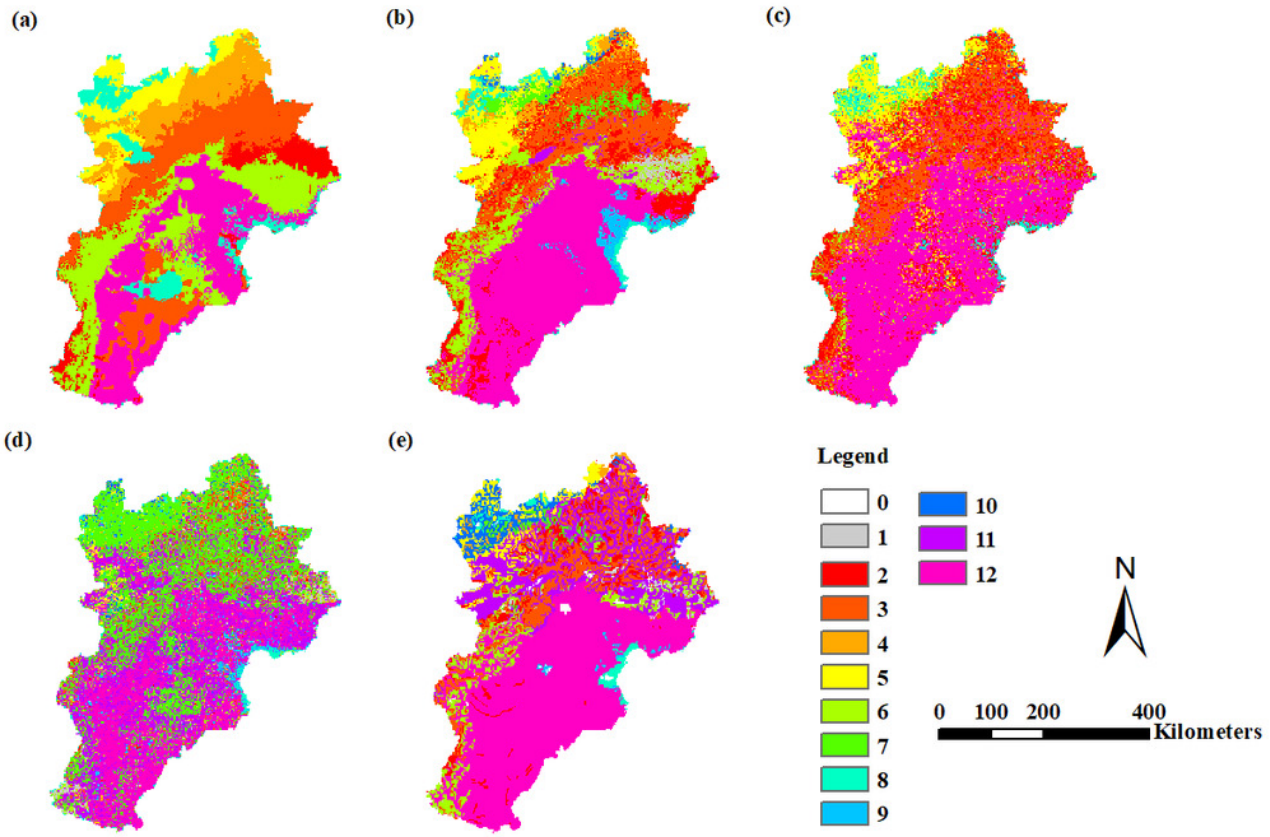

Figure 4

The modeling vegetation map of formations and sub-formations with highest accuracy by Decision tree model (a), Random forest model (b), Support vector machine (c), Maximum likelihood classification (d) and the Vegetation Map of the People's Republic of Ch

The legend represents formations and sub-formations shown in Table 1.