forests

MDPI

Article

Modeling Anthropogenic Fire Occurrence in the Boreal Forest of China Using Logistic Regression and Random Forests

Futao Guo [1,*], Lianjun Zhang [2], Sen Jin [3], Mulualem Tigabu [4], Zhangwen Su [1] and Wenhui Wang [1]

[1] College of Forestry, Fujian Agriculture and Forestry University, Fuzhou 350002, Fujian, China; fujianSZW@126.com (Z.S.); fafuwangedu@126.com (W.W.)
[2] Department of Forest and Natural Resources Management, College of Environmental Science and Forestry, State University of New York, Syracuse, NY 13210, USA; lizhang@esf.edu
[3] Faculty of Forestry, Northeast Forestry University, Harbin 150040, Heilongjiang, China; jinsen2005@126.com
[4] Southern Swedish Forest Research Centre, Swedish University of Agricultural Sciences, Box 49, Alnarp SE-230 52, Sweden; Mulualem.Tigabu@slu.se
* Correspondence: guofutao@126.com; Tel./Fax: +86-591-8378-0261

Academic Editors: Michael C. Stambaugh and Timothy A. Martin
Received: 15 June 2016; Accepted: 18 October 2016; Published: 25 October 2016

Abstract: Frequent and intense anthropogenic fires present meaningful challenges to forest management in the boreal forest of China. Understanding the underlying drivers of human-caused fire occurrence is crucial for making effective and scientifically-based forest fire management plans. In this study, we applied logistic regression (LR) and Random Forests (RF) to identify important biophysical and anthropogenic factors that help to explain the likelihood of anthropogenic fires in the Chinese boreal forest. Results showed that the anthropogenic fires were more likely to occur at areas close to railways and were significantly influenced by forest types. In addition, distance to settlement and distance to road were identified as important predictors for anthropogenic fire occurrence. The model comparison indicated that RF had greater ability than LR to predict forest fires caused by human activity in the Chinese boreal forest. High fire risk zones in the study area were identified based on RF, where we recommend increasing allocation of fire management resources.

Keywords: human-caused fire; driving factors; forest fire; Daxing'an Mountains; ROC curve

1. Introduction

Boreal forests (45°–70° north latitude) account for more than 25% of the world's forested areas [1], and provide vital natural and economic resources for northern circumpolar countries. In addition, they contain large belowground carbon pools in the form of peat [2,3]. Forest fire has been a major disturbance, influencing the energy flows and biogeochemical cycles in boreal forest ecosystems [4–6].

Human-caused or *anthropogenic* fires can be linked to a variety of human activities such as recreation (e.g., camping, hiking, hunting, etc.) and industry such as timber production or railway transportation. It is well-known that these activities play a critical role in fire occurrence in the boreal forest. In Ontario, Canada, on average, two-thirds of all forest fires were caused by humans over the 1976 to 1999 period [7]; while, in the Siberian boreal forest, human-caused fires are responsible for greater than 85% of the total forest fires [8,9]. Anthropogenic factors also dominate the fire regimes in the Chinese boreal forest [10]. Understanding the primary factors that influence human-caused fire occurrence is crucial and necessary for the allocation of fire prevention and suppression resources and forest management. Human activity and biophysical factors have been found to strongly influence anthropogenic fires. Korovin [11] found that most anthropogenic fires started close to roads. Thus, fire proximity to roads has been used as a variable in some fire prediction models [12]. Niklasson and Granström [13] and Wallenius et al. [14] indicated that expansion of human settlements and increased population density drove the fire occurrence in the boreal forest of northern Europe. Zumbrunnen et al. and Turco et al. [15,16] revealed the importance of weather factors such as temperature and precipitation on fire occurrence. Other factors such as topography (e.g., elevation and slope) and forest type were found to be meaningful drivers [9,17–20]. Socio-economic indicators, such as unemployment rate and population density, have also been linked to human-caused fire occurrence in many areas [21,22]. In the past decade, researchers have attempted to determine the driving factors and the probability of occurrence of human-caused fire in the Chinese boreal forest [23–25]. However, many quantitative analyses of forest fire drivers have paid less attention to the importance of socio-economic variables or human variables compared to climate-related and topographical variables. In this study we attempt to capture some of this complexity by considering the relationships among biophysical and human factors for human-caused fire occurrence.

Logistic regression (LR) is an approach commonly used to model the influence of different factors on fire occurrence (a binary response variable), and has been used in many studies [19,23,26,27]. However, logistic regression has its limitations. For example, the link function of logistic regression utilizes a linear function to regress the logged odds of fire occurrence to a set of independent or predictor variables. The result of LR modeling may also be affected by the flaws in the data such as outliers, multicollinearity among independent variables, and correlated observations. Because nonlinear and complex relationships often exist between biophysical and social variables, logistic regression may not always be sufficient and efficient [28].

Random Forests (RF) is an ensemble learning method based on classification and regression trees (CART). RF can select important variables and calculate the relative importance of each independent variable automatically no matter how many variables are used initially [29]. Additionally, RF has been demonstrated to have a high prediction accuracy and high tolerance to outliers and "noise" [30,31]. Due to the strengths of Random Forest, fire occurrence studies have begun using this approach in recent years [25,32].

In this study, we use both approaches (LR and RF) to evaluate the potential contribution of biophysical and anthropogenic factors to human-caused fire in the Chinese boreal forest. Each approach was applied to the fire data and the results of the two models were evaluated and compared. Furthermore, the maps of fire occurrence likelihood were created based on the results of the two approaches.

2. Materials and Methods

2.1. Study Site

China's boreal forest, located in the Daxing'an Mountains of northeastern China (50°10′–53°33′ N and 121°12′–127°00′ E), is the southernmost part of the global boreal forest biome. Its total area covers 8.46×10^6 ha (Figure 1). The dominant species is Dahurian larch (*Larix gmelinii* Rupr.), and is normally accompanied by white birch (*Betula platyphylla* Suk.) and Mongolian pine (*Pinus sylvestris L. var. mongolica* Litv.). The Daxing'an Mountains are located in the cold-temperate zone, with a mean annual temperature between −2 °C and 4 °C, and a range extending from −52.3 °C to 39.0 °C. The mean total annual precipitation is between 350–500 mm.

Boreal forest in this region was largely uninhabited until the construction of the first railway across the mountains in the early 20th century [25]. Before then, fire ignitions were assumed to have been caused primarily by lightning strikes [25,33]. After the introduction of the Reform and Open Policy by the Chinese Government in 1978, China has moved into a period of rapid development, leading to more frequent and intensified human-economic activities in the region of boreal forest. Today, this region has the largest average annual burned area in China and is generally exposed to extremely high fire risk due to the increases in forest-based economic activities. Between 1980 and 2005, there were more than 1000 forest fires, including more than 600 human-caused fires, and a total area of burned forest amounting to 1,300,000 ha [33]. In order to circumvent increased forest fire incidence and the costs of damage incurred by forest fires, a series of fire prevention and suppression policies have been issued since 1949 (the foundation of People's Republic of China), and revised after 1987 when the most serious forest fire of the century occurred in the Chinese boreal forest. In recent years, fires have become smaller (burned area), but occur more frequently and intensely than before [34].

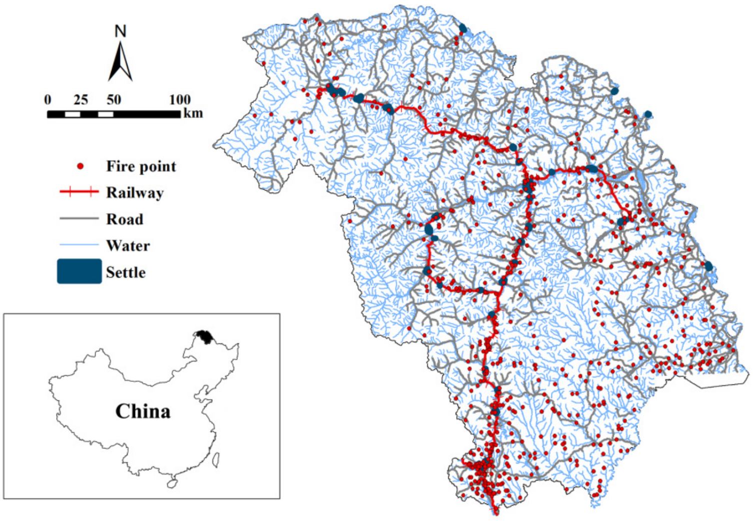

Figure 1. Map of study area—Daxing'an Mountain region of northeastern China, showing fire ignition points and human infrastructure.

2.2. Data Collection and Processing

Anthropogenic fire data for the Daxing'an Mountains from 1980 to 2005 were provided by the Forest Fire Prevention Office of Heilongjiang Forestry Bureau, P.R. China, which contained information on: fire location, size, cause, and date of occurrence. In this study, anthropogenic causes of forest fires included smoking, hunting, fireworks, escaped fire from locomotives and residents' homes, but not controlled prescribed burns and other intentional action taken by government or forest management agencies. The data provided by the office were in a geo-database format (ESRI data storage and management framework) and contained geographically referenced point locations of forest fires in the Daxing'an Mountains. Prior to 1990, the fire locations were determined by the fire chief, who identified each fire through a combined approach of fixed observation points in the forest and the Terrain and Forest Instruction Map (1:100,000). After 1990, the fire locations were recorded by Global Position System (GPS).

We created a binary variable (i.e., fire occurrence) for the logistic regression (LR) and Random Forests (RF) models. For each location of the observed fire points (620) the fire occurrence was coded 1 (representing "Yes"). Then, we randomly generated non-fire (i.e., control) points in the study area at a ratio of 1:1.5 as the fire ignition number [21,24], resulting in 905 control points where the fire occurrence was coded 0 (representing 'No'). We excluded control points located in water bodies or urban areas.

The independent or predictor variables consisted of five categories, including climate, vegetation, topography, infrastructure, and socio-economic factors. Details of these variables are provided in Table 1. The criteria for selecting the independent variables were based on previous studies of fire occurrence.

2.2.1. Climate Factors

Daily climate data were extracted from five national weather stations located in the study area. Daily climate data were provided by the China Meteorological Data and Sharing Network [35].

The dataset contains three climate factors, including daily mean temperature, daily precipitation, and daily mean relative humidity (Table 1). The corresponding daily climate factors for each fire and control point were retrieved under ArcGIS19.0 environment. The daily climate variables for those fire and control points were provided by the meteorological station that was identified as being closest to each point [10].

2.2.2. Vegetation

A digital vegetation map of China with 1 km resolution was downloaded from the Cold and Arid Regions Science Data Center, China [36]. We grouped polygons into the following five categories (Figure 2), with relative area of the Daxing'an Mountains reported in parenthesis: needle-leaf deciduous trees (30.6% cover of study area), broad-leaf deciduous trees (12.8%), needle-leaf evergreen trees (11.5%), broad-leaf deciduous shrub (7.45%), grass and agricultural crop (37.7%). The vegetation types for each fire and control point (i.e., non-fire) were extracted from the vegetation map layer using ArcGIS 10.2. We used the proportion of each vegetation type located in a fire or control point to develop the model.

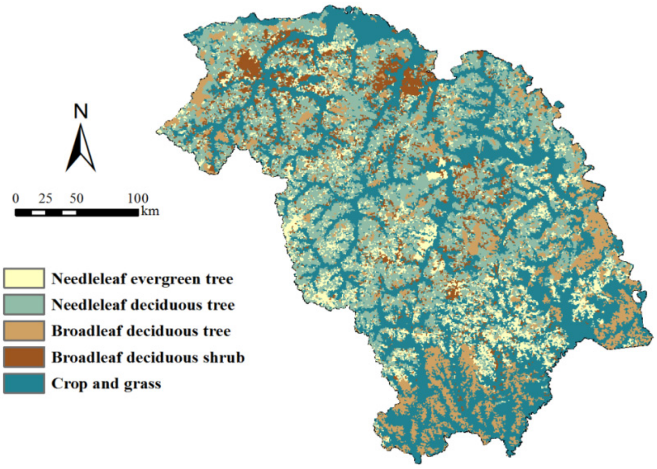
N
0 25 50 100
km
🟨 Needleleaf evergreen tree
🟩 Needleleaf deciduous tree
🟧 Broadleaf deciduous tree
🟫 Broadleaf deciduous shrub
🟦 Crop and grass

Figure 2. Distribution of forest vegetation types in the study area, Daxing'an Mountains, China.

2.2.3. Topography

Topographic features affect the spatial patterns of vegetation, plant assemblages, and relative flammability, in addition to influencing local climatic conditions. High resolution (25 m) digital elevation model (DEM) data were collected from the National Administration of Surveying, Mapping and Geo-information of China. The values associated with these DEMs for slope (° degree) and aspect were retrieved (Figure 3). We transformed the DEM-based aspect into cosine aspect (Cos_as) under ArcGIS 19.0 environment by using a trigonometric function so that Cos_as will be close to 1 when the aspect is generally northward, close to −1 when the aspect is southward, and close to 0 when the aspect is eastward or westward [35].

2.2.4. Infrastructure

In the past few decades, the influence of human infrastructure on wildfire has been widely studied [24,32,36–40]. However, certain types of human infrastructure were not considered in previous analyses of wildfire drivers in the Chinese boreal forest, such as the number of fire towers, number of inspection stations (the stations aim to inspect the potential fire ignition sources taken by people who want to get into the forest during the fire seasons), and length of burned line (the "burned line" is a measure for fire prevention, referring to the practice where forest managers burn ground forest fuel intentionally to decrease the risk of fire occurrence). In this study, we used a number of previously tested variables, but also included several unique, untested variables for fire prevention (Table 1).

Variables such as distance to the nearest railway, distance to the nearest road, and others were retrieved from a 1:250,000 Digital Line Graphic (DLG) map from the National Administration of Surveying, Mapping and Geo-information of China. The distribution of infrastructure is shown in Figure 1.

2.2.5. Socio-Economic Factors

Socio-economic factors included annual funding for forest fire prevention, population density, per capita GDP, and unemployment rates, which were collected from a statistical yearbook [38]. These variables have been used in other similar studies to represent trends surrounding potential changes in human activity, which may influence fire occurrence [36,41,42].

2.3. Models and Computing Procedures

2.3.1. Multicollinearity Test

High correlation between independent variables, namely multicollinearity, may exist in a linear regression model, which may distort the model estimation or interfere with accurate estimation. VIF (variance inflation factor) method was used to test for multicollinearity in this study, and variables with significant collinearity (VIF $\geq$ 10) were gradually removed from the models [24].

2.3.2. Models

(1) Logistic regression (LR). LR describes the relationship between a binary response variable (Y, coded as 0 (representing "No") and 1 (representing "Yes")) and one or more predictor variables (X) by means of a link function. It has been used for fire occurrence prediction and to examine the driving factors of fire occurrence in different regions of the world at various scales [19,40,41]. Logistic regression is commonly expressed as follows:

$$P\left(Y=1\right)=\frac{1}{1+e^{-X\beta}} \quad (1)$$

This model relates the probability of fire occurrence $P(Y = 1)$ with p predictor variables, which is a multiple linear regression model such that

$$\eta=X\beta=\beta_0+\sum_{j=1}^{p}\beta_j X_{ij}=\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots+\beta_p X_{ip} \quad (2)$$

where η is the link function in generalized linear models, and β_0–β_p are model coefficients to be estimated from data using maximum likelihood method. The logistic function (Equation (1)) lies between zero and one and takes on an f S-shaped curve.

(2) Random Forests (RF). RF is an ensemble learning technique for classification, regression, and other tasks. It operates by constructing a number of decision trees, where each tree is generated by bootstrapping samples, and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees [30]. Each decision tree uses two-thirds of the data to train RF, while the remaining one-third of the data (namely out of bag samples (OOB)) are retained for model validation [42]. In the modeling process, RF generates variable importance measures by comparing increases in OOB error when that variable is randomly permuted, while keeping all others unchanged [43,44]. RF is a nonparametric modeling algorithm which is robust to outliers and over-fitting and it also enables variable importance measures to be computed and compared to other regression techniques [30,45,46].

The original dataset was randomly divided into training (60%) and validation (40%) samples. This process was repeated five times, resulting in five random sub-samples of the data (each one with its own training and validation dataset). We applied LR to the training data of each sub-sample, creating five intermediate models. In order to validate the resultant intermediate models, each was tested with the validation samples (i.e., 40% of the original dataset). The final LR model was built using variables selected from the previous five intermediate models and applied to the whole dataset. The predictor variables that were statistically significant ($\alpha = 0.05$) in at least three of five intermediate models were included in the final LR model. A backward stepwise selection process was used during LR model fitting.

Because RF is generated via bootstrapping samples, it is not necessary to divide the complete dataset into training and validation parts. However, in order to use an analogous approach to LR, RF was also conducted using each sub-sample with the training datasets. Defining the number of variables to test at each split (*mtry*) and the number of trees to run (*ntree*) is required before running RF. Oliveira et al. [32] found that the increase in values of mtry would result in a higher predictive performance. Generally speaking, the parameter mtry was identified using the internal RF function *TuneRF*, this function computes the optimal number of variables starting from the default ($mtry = \sqrt{\text{total number of variables}}$ for classification) and it searches below and above this threshold for the value with the minimum OOB error rate [47]. The *ntree* parameter was set to 1000 in order to obtain stable results. The most relevant independent variables were selected based on their importance in each sub-sample. Similar to the LR model, the variables that were most relevant in at least three out of five intermediate models were then included in the final model, which was fitted using the complete data set.

An alternative method of quantifying predictive ability for both LR and RF models is receiver operating characteristic (ROC) analysis [48]. The ROC curve was obtained by plotting sensitivity versus specificity for various probability thresholds. The area under the curve (AUC) is also often used to evaluate performance [49]. An AUC of 0.5 indicates no discrimination, 0.5–0.69 poor discrimination, 0.7–0.79 reasonable discrimination, 0.8–0.9 excellent discrimination [50]. In other words, higher AUC indicates better performance of model fitting.

In addition, we determined a probability threshold (i.e., the cut-off value) based on Yueden index [24], which has previously been used to determine the best cut-off values in logistic regression for predicting wildfire occurrence [21,24,51]. The calculation of Yueden index (the best cut-off value) is based on the sensitivity and specificity of ROC "sensitivity + specificity − 1". If the predicted fire occurrence was at or above the cut-off value, the occurrence of fire ignition was considered to have occurred. Otherwise, it was registered as no fire [24,52].

As per mapping the likelihood of fire occurrence, maps showing the fire occurrence likelihood were created using the Kriging method in an ArcGIS environment, and were based on the predicted fire occurrence by both LR and RF models using the whole dataset.

3. Results

3.1. Test for Multicollinearity

Two independent variables, length of road construction and number of fire towers, were removed since the multicollinearity was identified based on the VIF test. The remaining 17 independent variables were selected for logistic regression model fitting.

3.2. Identification of Driving Factors by LR and RF Models

Five intermediate models were created based on sub-samples of the dataset. Four variables, including forest type (Forest_type), distance to the railways (Dis_railway), distance to the roads (Dis_road), and distance to the settlements (Dis_sett) were found to be significant in at least one of five sub samples, but only forest type and distance to the railways were significant in more than three sub samples (Table 2). Thus, these two variables were included in subsequent analyses of the final model fitted with the complete dataset. Table 2 also shows the parameter estimation of the final LR model. Forest type and distance to the railways included in the final model were both significant at the significance level α = 0.01. Thus, the final LR model for predicting the probability of fire occurrence $P(Y = 1)$ is

$$P\left(Y=1\right)=\frac{1}{1+e^{-\left(-0.166\,+\,2.419\cdot \text{Forest\_type}\,-\,0.00005\cdot \text{Dis\_railway}\right)}} \tag{3}$$

The overall prediction accuracy of the final LR model using the whole dataset was 60.8%.

For the RF models, the variable importance plots for the five sub-samples were obtained according to the minimum OOB error principle (Figure 4). We utilized the variables that were significant in at least three of the five intermediate models to fit the complete dataset using the RF model. The most important climate variables in the final model according to the values of % IncMSE are shown in Table 3, including distance to the railways, distance to the settlements, forest type, and distance to the roads (in descending order).

Table 2. Comparison of prediction accuracy and goodness of fit between LR and RF models.

Sample	Model Type	Cut-off	Prediction Accuracy (%)	
			Training Data	Validation
Sample1	LR/RF	0.61/0.59	63.5/71.0	53.8/69.6
Sample2	LR/RF	0.61/0.59	60.8/68.3	64.5/66.8
Sample3	LR/RF	0.58/0.65	60.3/69.4	59.2/68.5
Sample4	LR/RF	0.62/0.62	61.5/72.6	53.6/71.3
Sample5	LR/RF	0.59/0.57	58.8/70.3	58.1/67.4
Complete dataset	LR/RF	0.60/0.59	60.8/70.1	

Note: LR, logistic regression; RF, Random Forest; Cut-off values were used to determine the prediction accuracy for each intermediate and final model. The complete dataset was not divided into training and validation subsets.

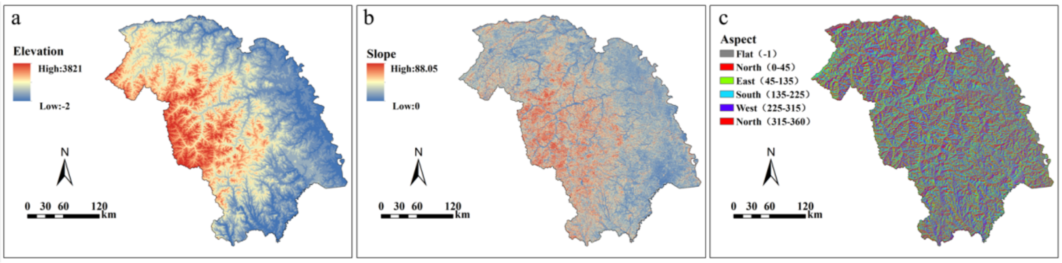

Figure 3. Elevation (**a**), slope (**b**), and aspect (**c**) throughout the study area.

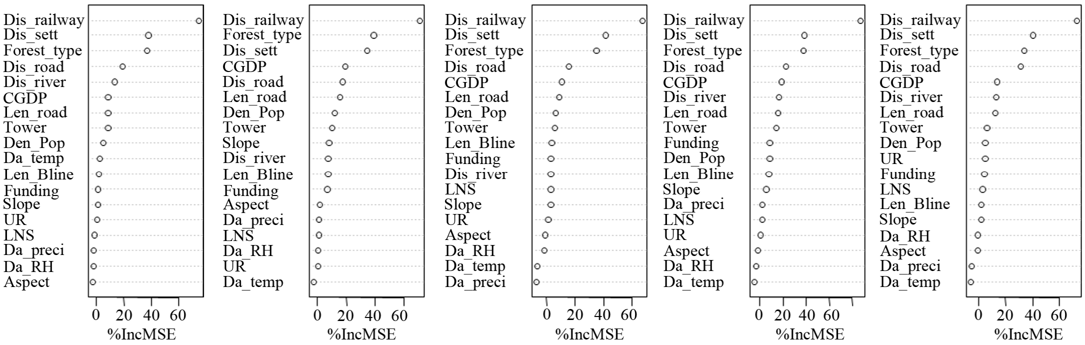

Figure 4. Variable importance measures from Random Forest sub-samples based on Mean Decrease Accuracy (*X*-axis), which quantifies the importance of a variable by measuring the change in prediction accuracy when the values of the variable are randomly permuted compared with the original observations. The abbreviated variable names are the same as in Table 1.

3.3. Comparison of LR and RF Performance

We calculated the prediction accuracy and generated ROC curve of each sub-sample and complete dataset in order to test and compare the predictive ability of LR and RF. Table 4 showed that the correct prediction rate of LR ranged from 53.6%–64.5% for the intermediate models, and 60.8% for the final LR model using the whole dataset (Table 4). In contrast, the prediction accuracy of RF for sub-samples was 66.8%–72.6%, and 70.1% for the final RF model using the whole dataset. The ROC curves (Figure 5) indicated that RF performed better in intermediate models using the sub-sample dataset, as well as the final model using the complete dataset in terms of AUC values.

Table 3. Variables included in the final model using Random Forests, in descending order of importance based on mean decrease in accuracy from the complete data set.

Variable	Mean Decrease in Accuracy (%IncMSE)
Dis_railway	87.6660
Dis_sett	66.2240
Forest_type	61.3847
Dis_road	36.4337

Table 4. Variables identified by intermediate models using logistic regression and parameter estimation using the selected variable.

Variables Identified by Intermediate Models				Parameter Estimation			
Variable	*p*-Value (min)	*p*-Value (max)	Significant Samples	Coefficients	Standard Error	Wald Test	*p*-Value
Forest_type	<0.0001	0.007	5	2.419	0.550	19.357	<0.0001
Dis_railway	0.008	0.073	3	−0.00005	0.00002	21.563	0.008
Dis_road	0.005	0.005	1	–	–	–	–
Dis_sett	0.003	0.003	1	–	–	–	-

Note: *p*-value (min and max) in the table represent the minimum and maximum significant level of each variable in the intermediate models; significant samples represents the total number that the variable is tested as significant in the five intermediate models. The four columns on the right of the table show the parameter estimation of selected important variables in the final model that was fitted based on the complete dataset. The abbreviated variable names are the same as in Table 1.

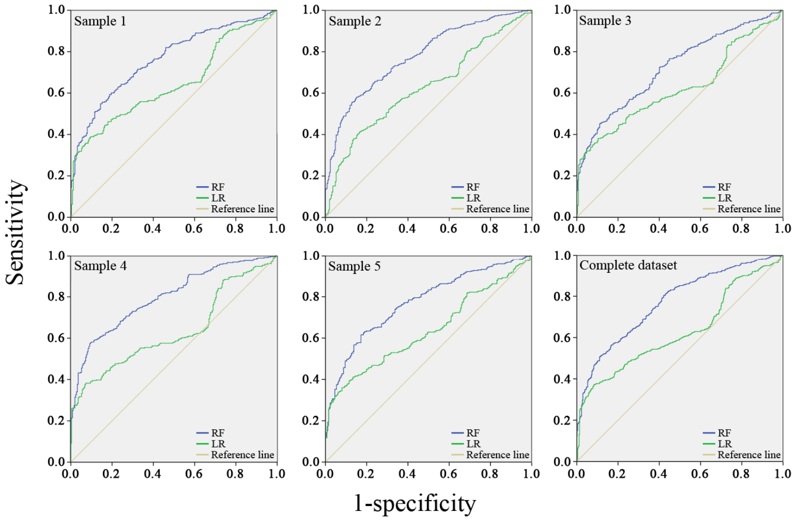

Figure 5. ROC (receiver operating characteristic) curves of each sub-sample and complete dataset using Random Forests (RF) and logistic regression (LR) models. The upper curve has a bigger area under curve (AUC) meaning higher predictive ability.

Table 1. Independent or predictor variables included in forest fire model development for Daxing'an Mountains.

Variable Type	Variable Name	Code	Resolution/Scale	Description	Source/Reference
Climatic	Daily precipitation	Da_preci	Daily/0.01	The corresponding daily climate factors of each fire point and control point based on five national weather stations	[52]
	Daily mean relative humidity	Da_RH			
	Daily mean temperature	Da_temp			
Topographic	Slope	Slope	Raster/25 m	The slope of each fire point and control extracted from a raster map of study area	[53]
	Aspect	Aspect		Generate aspect from ArcGIS as a continuous variable and transform. Aspect was transformed into cosine aspect (Cos_as) by using a trigonometric function	
Vegetation	Forest type	Forest_type	Raster/1 km	Proportion of each forest type in the study area	[54]
Infrastructure	Distance to nearest railway	Dis_railway	Vector/1:250,000	The straight distance between a fire point or a control point and the nearest railway	[53]
	Distance to nearest river	Dis_river		The straight distance between a fire point or a control point and the nearest river	
	Distance to nearest road	Dis_road		The straight distance between a fire point or a control point and the nearest road	
	Distance to nearest settlement	Dis_sett		The straight distance between a fire point or a control point and the nearest settlement	
	Number of fire towers	Tower	Yearly/0.1	The number of fire towers that were used to monitor fire occurrence	[55]
	Number of inspection stations	LNS		The number of inspection stations that were used to inspect the potential fire source with people who will enter the mountains during the fire season	
	Length of burned line	Len_Bline		The length of burned line for fire prevention	
	Length of road construction	Len_road		The length of road for fire prevention	
Socio-economic	Per Capita GDP	CGDP	Yearly/0.01	Per capita GDP of the study area	[56]
	Unemployment rate	UR		The unemployment rate of the study area	
	Population density	Den_Pop		The annual population density of the study area	
	Funding	Funding		Annual funding for forest fire prevention	[55]

Note that the code is a short variable name used in the models.

According to the LR final model, high fire risk spots were identified at the northern, eastern, and southern study area, respectively (Figure 6). The maps of likelihood of fire occurrence produced by the RF final model, however, provided more defined fire risk for the study area. The high likelihood of fire occurrence was concentrated in the middle and southern portions of the study area (Figure 6).

We also conducted residual analysis to compare the LR and RF final models based on the whole dataset. Figure 7 demonstrated that the RF final model had the best fit (i.e., overall smaller residuals across the study area). In contrast, the overall residual of the LR final model was higher than that of RF and spatially uneven compared to RF. The maximum (0.55) and minimum (−0.55) residuals of RF were both lower than that of LR (0.55 and −0.65, respectively). The positive (under-prediction) and negative (over-prediction) residuals of LR model were also clustered within the study area. The positive residuals were mainly located in the middle and southern portions of the study area, while the negative residuals were concentrated on the northern and northeastern portions of the study area. In comparison, the RF final model had relatively smaller residuals across the entire study area, and a small negative residual cluster was distributed in the north of the study area, and two small positive residual clusters were located in the mid- and southern portions of the study area (Figure 7).

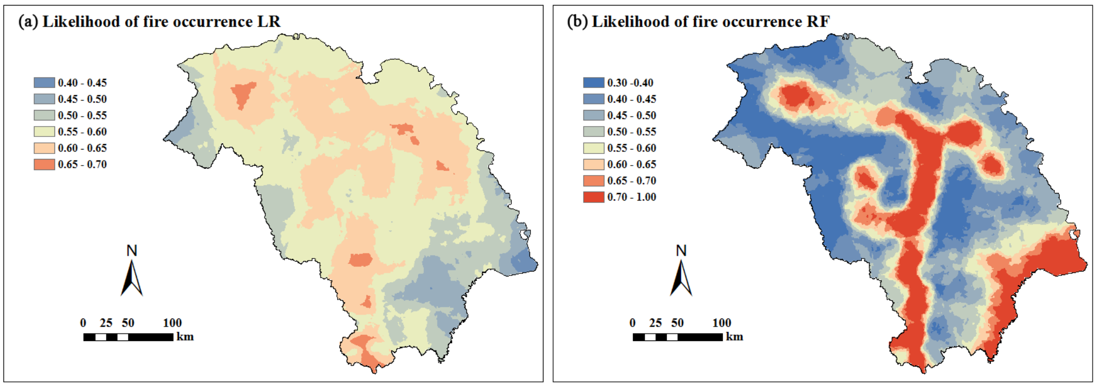

Figure 6. Likelihood of fire occurrence created using the Kriging method with ArcGIS based on predicted values using Logistic Regression (**a**) and Random Forests (**b**) models.

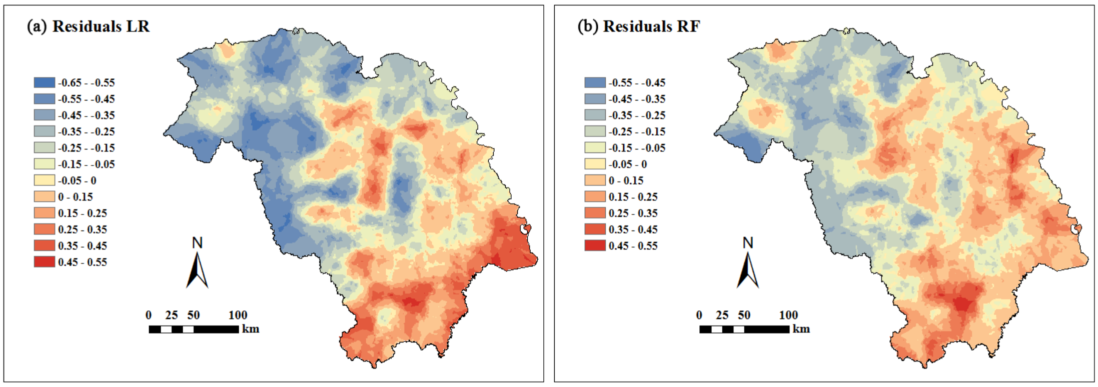

Figure 7. Spatial distribution of model residuals of logistic regression (**a**) and Random Forests (**b**) models.

4. Discussion

In this study, forest type and distance to the nearest railway were identified as the most important driving factors in both LR and RF models, while distance to the settlement and distance to the roads were identified as useful predictors in the RF models only.

Forest type reflects the fuel conditions, which significantly influences the fire ignition of the study area. Other studies have revealed the importance of transportation corridors to forest fire occurrence, such as railways and roads [20,24,51]. In this study, both railways and roads were also identified as important drivers on human-caused fires by the RF models. Distance to the settlement is typically described as wildland urban interface (WUI), which appeared to be another driver of anthropogenic fire in the Chinese boreal forest.

Socio-economic factors, such as GDP, unemployment, and population density, did not appear to play a crucial role on local human-caused fire occurrence in either LR or RF models, which is consistent with similar studies [23,24]. One possible explanation is that economic and social development in the Chinese boreal forest has been relatively slow, which may not have resulted in significant impacts on anthropogenic forest fire during the past few decades. An additional influence may be China's birth control policy (i.e., one child per family), which has tended to stabilize population increases, and, therefore, would not significantly influence anthropogenic fire occurrence more broadly [25]. Similar to socio-economic factors, the importance of climate factors were not identified by the models, demonstrating that human, topographic, and fuel factors dominate fire occurrence in the study area.

The RF final model included predictor variables that were not present in the final LR model, such as distance to the roads and distance to the settlement. The reason for the difference is possibly attributable to the variable selection method used in the LR model. We used a backward stepwise approach to identify which variables were significant. However, Harrell et al. [29] suggested that stepwise variable selection based on the significance level as the criterion for entering a variable, but this would not take into account the problem of multiple comparisons and may influence the selection of real-power variables.

In this study, the RF model seemed to perform better than the LR model with regards to fire prediction accuracy since the RF model had higher predictive capacity for human-caused fire occurrence in the Chinese boreal forest. According to anthropogenic fire likelihood maps based on the RF final model, southern, southeastern, and middle regions of the study area were identified as having higher fire risks. In these high fire risk regions, efficacy and efficiency of fire prevention strategies and use of resources (e.g., fire towers, inspection stations, fire patrols, etc.) might be improved by focusing these in low elevation zones, along railways, and around residential areas during the fire season.

5. Conclusions

We applied logistic regression and Random Forests to identify which biophysical and human activity factors were important drivers of anthropogenic fire in the Chinese boreal forest. Both methods indicated that forest type and distance to the railways significantly influenced anthropogenic fires. In addition, predictor variables such as distance to the settlement and distance to the roads were also identified by one model as useful factors affecting anthropogenic fire occurrence. Our results revealed that anthropogenic fires were more likely to occur close to infrastructures such as railways, roads, and settlements, and were also significantly influenced by forest types.

Socio-economic factors such as GDP, unemployment, and population density were not identified as important driving factors on anthropogenic fires in the Chinese boreal forest, which may be due to the relatively slower economic development and population increase.

Compared to logistic regression, Random Forest had an increased ability to predict forest fires caused by human activities in the Chinese boreal forest. According to the spatial distribution of fire occurrence likelihood computed by the RF final model, three "hot spots" were identified in the southern, southeastern, and middle regions of the study area. Our findings provide an important form of guidance for local forest fire management in terms of considering fire resource allocation (e.g., fire towers, inspection stations, fire patrols, etc.), which could improve the efficiency of forest fire management in this region of China.

Acknowledgments: This research was financially funded by the National Natural Science Foundation of China (Grant No. 31400552), the Fujian Agriculture and Forestry University Funds for Distinguished Young Scholar (Grant No. xjq201613), and Natural Science Foundation of Fujian Province (Grant No. 2015J05049).

Author Contributions: Futao Guo, Zhangwen Su, Wenhui Wang conceived and designed the experiment, collected and analyzed the data, and wrote the original draft. Lianjun Zhang, Sen Jin, Mulualem Tigabu critically reviewed and edited the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Intergovernmental Panel on Climate Change (IPCC). *Land Use, Land-Use Change, and Forestry*; Cambridge University Press: Cambridge, UK, 2000; p. 4.

2. Gorham, E. Northern peatlands: Role in the carbon cycle and probable responses to climate warming. *Ecol. Appl.* **1991**, *1*, 182–195. [CrossRef] [PubMed]
3. Zimov, S.A.; Davidov, S.P.; Zimova, G.M.; Davidova, A.I.; Chapin, F.S.; Chapin, M.C.; Reynolds, J.F. Contribution of disturbance to increasing seasonal amplitude of atmospheric CO_2. *Science* **1999**, *284*, 1973–1976. [CrossRef] [PubMed]
4. Weber, M.G.; Flannigan, M.D. Canadian boreal forest ecosystem structure and function in a changing climate: Impacts on fire regimes. *Environ. Rev.* **1997**, *5*, 145–166. [CrossRef]
5. Stocks, B.J. Global warming and forest-fires in Canada. *For. Chron.* **1993**, *69*, 290–293. [CrossRef]
6. Shvidenko, A.Z.; Goldammer, J.G. Fire situation in Russia. *Int. For. Fire News* **2001**, *24*, 41–59.
7. Wotto, B.M.; Martell, D.L.; Logan, K.A. Climate change and people-caused forest fire occurrence in Ontario. *Clim. Chang.* **2003**, *60*, 275–295. [CrossRef]
8. Shvidenko, A.Z.; Nilsson, S. Fire and the carbon budget of Russian forests. In *Fire, Climate Change, and Carbon Cycling in the Boreal Forest*; Kasischke, E.S., Stocks, B.J., Eds.; Springer: New York, NY, USA, 2000; pp. 289–311.
9. Mollicone, D.; Eva, H.D.; Achard, F. Human role in Russian wild fires. *Nature* **2006**, *440*, 436–437. [CrossRef] [PubMed]
10. Guo, F.; Innes, L.J.; Wang, G.; Ma, X.; Sun, L.; Hu, H.; Su, Z. Historic distribution and driving factors of human-caused fires in the Chinese boreal forest between 1972 and 2005. *J. Plant Ecol.* **2015**, *8*, 480–490. [CrossRef]
11. Korovin, G.N. Analysis of the distribution of forest fires in Russia. In *Fire in Ecosystems of Boreal Eurasia*; Goldammer, J.G., Furyaev, V.V., Eds.; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1996; pp. 112–128.
12. Cardoso, M.F.; Hurtt, G.C.; Moore, B.; Nobre, C.A.; Prins, E.M. Projecting future fire activity in Amazonia. *Glob. Chang. Biol.* **2003**, *9*, 656–669. [CrossRef]
13. Niklasson, M.; Granström, A. Numbers and sizes of fires: Long-term spatially explicit fire history in a Swedish boreal landscape. *Ecology* **2000**, *81*, 1484–1499. [CrossRef]
14. Wallenius, T.H.; Kuuluvainen, T.; Vanha-Majamaa, I. Fire history in relation to site type and vegetation in Vienansalo wilderness in eastern Fennoscandia, Russia. *Can. J. For. Res.* **2004**, *34*, 1400–1409. [CrossRef]
15. Zumbrunnen, T.; Pezzatti, G.B.; Mene'ndezd, P.; Bugmann, H.; Bürgi, M.; Conedera, M. Weather and human impacts on forest fires: 100 years of fire history in two climatic regions of Switzerland. *For. Ecol. Manag.* **2011**, *261*, 2188–2199. [CrossRef]
16. Turco, M.; Llasat, M.C.; Hardenberg, J.; Provenzale, A. Impact of climate variability on summer fires in a Mediterranean environment (northeastern Iberian Peninsula). *Clim. Chang.* **2013**, *116*, 665–678. [CrossRef]
17. Syphard, A.D.; Radeloff, V.C.; Keely, J.E.; Hawbaker, R.J.; Clayton, M.K.; Stewart, S.I.; Hammer, R.B. Human influence on California Fire Regimes. *Ecol. Appl.* **2007**, *17*, 1388–1402. [CrossRef] [PubMed]
18. Romero-Calcerrada, R.; Barrio-Parra, F.J.; Millington, D.A.; Novillo, C.J. Spatial modelling of socioeconomic data to understand patterns of human-caused wildfire ignition risk in the SW of Madrid (central Spain). *Ecol. Model.* **2010**, *221*, 34–45. [CrossRef]
19. Martínez, J.; Vega-García, C.; Chuvieco, E. Human-caused wildfire risk rating for prevention planning in Spain. *J. Environ. Manag.* **2009**, *90*, 1241–1252. [CrossRef] [PubMed]
20. Romero-Calcerrada, R.; Novillo, C.J.; Millington, J.D.A.; Gomez-Jimenez, I. GIS analysis of spatial patterns of human-caused wildfire ignition risk in the SW of Madrid (Central Spain). *Landsc. Ecol.* **2008**, *23*, 341–354. [CrossRef]
21. Catry, F.X.; Rego, F.C.; Bacao, F.L.; Moreira, F. Modeling and mapping wildfire ignition risk in Portugal. *Int. J. Wildland Fire* **2009**, *18*, 921–931. [CrossRef]
22. Chuvieco, E.; Aguado, I.; Yebra, M.; Nieto, H.; Salas, P.J.; Martín, M.P. Development of a framework for fire risk assessment using remote sensing and geographic information system technologies. *Ecol. Model.* **2010**, *221*, 46–58. [CrossRef]
23. Liu, Z.; Yang, J.; Chang, Y.; Weisberg, P.J.; He, H.S. Spatial patterns and drivers of fire occurrence and its future trend under climate change in a boreal forest of Northeast China. *Glob. Chang. Biol.* **2012**, *18*, 2041–2056. [CrossRef]

24. Chang, Y.; Zhu, Z.L.; Bu, R.C.; Chen, H.G.; Feng, Y.T.; Li, Y.H.; Hu, Y.M.; Wang, Z.C. Predicting fire occurrence patterns with logistic regression in Heilongjiang Province, China. *Landsc. Ecol.* **2013**, *28*, 1989–2004. [CrossRef]
25. Wu, Z.E.; He, H.S.; Yang, J.; Liu, Z.H.; Liang, Y. Relative effects of climatic and local factors on fire occurrence in boreal forest landscapes of northeastern China. *Sci. Total Environ.* **2014**, *493*, 472–480. [CrossRef] [PubMed]
26. Cardille, J.A.; Ventura, S.J.; Turner, M.G. Environmental and social factors influencing wildfires in the Upper Midwest, United States. *Ecol. Appl.* **2001**, *11*, 111–127. [CrossRef]
27. Prasad, V.K.; Badarinathb, K.V.S.; Eaturu, A. Biophysical and anthropogenic controls of forest fires in the Deccan Plateau, India. *J. Environ. Manag.* **2008**, *86*, 1–13. [CrossRef] [PubMed]
28. Schoenberg, F.P.; Peng, R.; Huang, Z.J.; Rundel, P. Detection of non-linearities in the dependence of burn area on fuel age and climatic variables. *Int. J. Wildland Fire* **2003**, *12*, 1–6. [CrossRef]
29. Cutler, D.R.; Edwards, T.C.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J.; Lawler, J.J. Random forests for classification in ecology. *Ecology* **2007**, *88*, 2783–2792. [CrossRef] [PubMed]
30. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
31. Archibald, S.; Roy, D.P.; van Wilgen, B.W.; Scholes, R.J. What limits fire? An examination of drivers of burnt area in Southern Africa. *Glob. Chang. Biol.* **2009**, *15*, 613–630. [CrossRef]
32. Oliveira, S.; Oehler, F.; San-Miguel-Ayanz, J.; Camia, A.; Pereira, J.M.C. Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. *For. Ecol. Manag.* **2012**, *275*, 117–129. [CrossRef]
33. Xu, H.C.; Li, Z.D.; Qiu, Y. Fire disturbance history in virgin forest in northern region of Daxinganling Mountatins. *Acta Ecol. Sin.* **1997**, *17*, 337–343.
34. Chang, Y.; He, H.S.; Bishop, I.; Hu, Y.M.; Bu, R.C.; Xu, C.G.; Li, X. Long-term forest landscape responses to fire exclusion in the Great Xing'an Mountains, China. *Int. J. Wildland Fire* **2007**, *16*, 34–44. [CrossRef]
35. Zhang, H.J.; Qi, P.C.; Guo, G.M. Improvement of fire danger modelling with geographically weighted logistic model. *Int. J. Wildland Fire* **2014**, *23*, 1130–1146. [CrossRef]
36. Oliveira, S.; Pereira, J.M.C.; San-Miguel-Ayanz, J.; Lourenço, L. Exploring the spatial patterns of fire density in Southern Europe using Geographically Weighted Regression. *Appl. Geogr.* **2014**, *51*, 143–157. [CrossRef]
37. Mundo, I.A.; Thorsten, W.; Rajapandian, K.; Thomas, K. Environmental drivers and spatial dependency in wildfire ignition patterns of northwestern Patagonia. *J. Environ. Manag.* **2013**, *123*, 77–87. [CrossRef] [PubMed]
38. Heilongjiang Statistics Bureau. *The Road of Revitalization: Thirty Years of Reform, Heilongjiang, 2009*; China Statistics Press: Beijing, China, 2009.
39. Maingi, J.K.; Henry, M.C. Factors influencing wildfire occurrence and distribution in eastern Kentucky, USA. *Int. J. Wildland Fire* **2007**, *16*, 23–33. [CrossRef]
40. Martell, D.L.; Otukol, S.; Stocks, B.J. A logistic model for predicting daily people-caused forest fire occurrence in Ontario. *Can. J. For. Res.* **1987**, *17*, 394–401. [CrossRef]
41. Vega-Garcia, C.; Woodard, T.; Adamowicz, W.L.; Lee, B. A logit model for predicting the daily occurence of human caused forest fires. *Int. J. Wildland Fire* **1995**, *5*, 101–111. [CrossRef]
42. Duro, D.C.; Franklin, S.E.; Dube, M.G. Multi-scale object-based image analysis and feature selection of multi-sensor earth observation imagery using random forests. *Int. J. Remote Sens.* **2012**, *33*, 4502–4526. [CrossRef]
43. Marston, C.G.; Danson, F.M.; Armitage, R.P.; Giraudoux, P.P.; Pleydell, D.R.J.; Wang, Q.; Qui, J.M.; Craig, P.S. A random forest approach for predicting the presence of Echinococcus multilocularis intermediate host *Ochotona* spp. presence in relation to landscape characteristics in western China. *Appl. Geogr.* **2014**, *55*, 176–183. [CrossRef] [PubMed]
44. Abdel-Rahman, E.M.; Ahmed, F.B.; Ismail, R. Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1Hyperion hyperspectral data. *Int. J. Remote Sens.* **2013**, *34*, 712–728. [CrossRef]
45. Perdiguero-Alonso, D.; Montero, F.E.; Kostadinova, A.; Raga, J.A.; Barrett, J. Random forests, a novel approach for discrimination of fish populations using parasites as biological tags. *Int. J. Parasitol.* **2008**, *38*, 1425–1434. [CrossRef] [PubMed]
46. Gromping, U. Variable importance assessment in regression: Linear regression versus random forest. *Am. Stat.* **2009**, *63*, 308–319. [CrossRef]

47. Genuer, R.; Poggi, J.M.; Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236. [CrossRef]
48. Swets, J.A. Measuring the accuracy of diagnostic systems. *Science* **1988**, *240*, 1285–1293. [CrossRef] [PubMed]
49. Jimenez-Valverde, A. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modeling. *Glob. Ecol. Biogeogr.* **2012**, *21*, 498–507. [CrossRef]
50. Del Hoyo, V.L.; Isabel, M.P.M.; Vega, F.J.M. Logistic regression models for human-caused wildfire risk estimation: Analyzing the effect of the spatial accuracy in fire occurrence data. *Eur. J. For. Res.* **2011**, *130*, 983–996. [CrossRef]
51. Stephens, S.L. Forest fire causes and extent on United States Forest Service lands. *Int. J. Wildland Fire* **2005**, *14*, 213–222. [CrossRef]
52. Daily Climate Data Set of China International Exchange Station, China Meteorological Data and Sharing Network. Available online: http://data.cma.cn/site/index.html (accessed on 15 May 2016).
53. Geographic Information Resources Service. National Administration of Surveying, Mapping and Geo-information of China. 2002. Available online: http://www.webmap.cn/main.do?method=index (accessed on 15 May 2016).
54. Ran, Y.H.; Li, X. Plant Functional Types Map in China. Cold and Arid Regions Science Data Center at Lanzhou, 2011. Available online: http://westdc.westgis.ac.cn/ (accessed on 15 May 2016).
55. Anonymous. *The Local Chronicles of Forest Fire Prevention of Daxing'an Mountains*; Heilongjiang People's Press: Haerbin, China, 2005.
56. Anonymous. *The Road of Revitalization-Thirty Years of Reform*; Statistics Press: Beijing, China, 2009.