A Prediction Model for the Outbreak Date of Spring Pollen Allergy in Beijing Based on Satellite-Derived Phenological Characteristics of Vegetation Greenness

Xinyi Yang 1,2, Wenquan Zhu 1,2,* and Cenliang Zhao 1,2

1 State Key Laboratory of Remote Sensing Science, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China
2 Beijing Engineering Research Center for Global Land Remote Sensing Products, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China
* Correspondence: zhuwq75@bnu.edu.cn

Abstract: Pollen allergies have a serious impact on people’s physical and mental health. Accurate and efficient prediction of the outbreak date of pollen allergies plays an important role in the conservation of people sensitive to allergenic pollen. It is a frontier research to combine new social media data and satellite data to develop a model to forecast the outbreak date of pollen allergies. This study extracted the real outbreak dates of spring pollen allergies from Sina Weibo records from 2011 to 2021 in Beijing and calculated five vegetation indices of three vegetation types as phenological characteristics within the 30 days before the average outbreak date. The sensitivity coefficients and correlation coefficients were used to screen the phenological characteristics that best reflected the outbreak date of spring pollen allergy. Based on the best characteristic, two kinds of prediction models for the outbreak date of spring pollen allergy in Beijing were established (the linear fit prediction model and the cumulative linear fit prediction model), and the root mean square error (RMSE) was calculated as the prediction accuracy. The results showed that (1) the date of EVI2 (2-band enhanced vegetation index) in evergreen forest first reaching 0.138 can best reflect the outbreak date of pollen allergies in spring, and (2) the cumulative linear fit prediction model based on EVI2 in evergreen forests can obtain a high accuracy with an average RMSE of 3.6 days, which can predict the outbreak date of spring pollen allergies 30 days in advance. Compared with the existing indirect prediction models (which predict pollen concentrations rather than pollen allergies), this model provides a new direct way to predict pollen allergy outbreaks by using only remote sensing time-series data before pollen allergy outbreaks. The new prediction model also has better representativeness and operability and is capable of assisting public health management.

Keywords: pollen allergy; vegetation phenology; satellite-derived greenness characteristics; vegetation index; prediction model

1. Introduction

Pollen allergy is a disease that causes patients to present with IgE-induced (immunoglobulin E) allergy symptoms after inhalation or exposure to pollen allergens [1], such as allergic rhinitis, allergic conjunctivitis, asthma, urticaria, and allergic dermatitis. In Europe, more than 15% of the population suffers from pollen allergies, and the proportion of pollen allergy sufferers is much higher in urban areas [2]. In China, the probability of diseases caused by pollen allergies is 0.5–1.0%, reaching approximately 5% in densely populated areas [3], and the rate of incidence is increasing continuously [4]. For pollen-allergic patients, the most direct way of relieving the symptoms of pollen allergy is to suggest the use of protective gear during the pollen allergy period. Hence, effectively predicting the outbreak date of pollen allergies can help to inform patients for when protection is necessary.
Airborne pollen is an important cause of pollen allergies [5], and the outbreak date of pollen is closely related to vegetation phenology. Remote sensing data have been widely used in vegetation phenology monitoring [6,7]. Therefore, using remote sensing data to reveal the outbreak date of pollen allergies and the greenness characteristics before the outbreak date can facilitate the prevention of pollen allergy diseases.

Current related studies on pollen allergy monitoring based on remote sensing data mainly focus on detecting pollen sources [8,9], extracting the flowering date [10,11], and identifying spectral features of pollen [12–14]. Satellite-based monitoring of the flowering date is difficult due to the short flowering period (5–30 days) and the relatively weak spectral signals of flowers at large scales. Satellite data with coarse spatial resolution (e.g., Landsat and MODIS) have fewer applications for monitoring flowering dates due to their low ability to distinguish optical signals of flowering. However, the flowering date is closely related to the phenological characteristics of vegetation greenness [15]. Therefore, the satellite data with moderate and low spatial resolution are able to indirectly reflect the flowering date through the earlier greenness characteristics of vegetation despite its inability to directly detect the spectral signals of flowering, which provides a possibility to monitor the outbreak date of pollen allergy.

The flowering period and pollen concentration are closely related to the symptoms of pollen allergy sufferers [16], so the current prediction models for pollen allergy outbreaks are mainly established to forecast the start date of flowering and the concentration of airborne pollen. Høgda et al. [17] predicted the start of the pollen season of Nordic birch based on the relationship between the normalized difference vegetation index (NDVI) and the phenology records of sites. Karlsen et al. [10] produced a map to characterize the onset of the birch pollen season utilizing NDVI satellite data. Khwarahm et al. [18] developed a technique to estimate the flowering phenophase of birch and grass from MERIS terrestrial chlorophyll index (MTCI) time-series data. Since the start date of the pollen season was recorded by monitoring the concentration of airborne pollen at sites, some studies started to predict the outbreak date of pollen allergies through the empirical relationship between pollen concentration and meteorological parameters [19]. For instance, He et al. [20] established a statistical model for pollen concentration prediction in Beijing combined with meteorological data. Iglesias et al. [21] developed a pollen concentration prediction model of sycamores in northwestern Spain based on temperature data. Myszkowska et al. [22] constructed a pollen concentration prediction model based on the relationship between the pollen concentration of multiple vegetation types and meteorological parameters in southern Poland. However, pollen monitors can only reflect pollen concentrations at the site scale because of their limited spatial coverage. Predicting pollen concentration based on the empirical relationship between pollen concentration and meteorological parameters tends to ignore the effects of other environmental factors on pollen concentration [23]. Therefore, some studies have developed pollen concentration prediction models based on remote sensing data, such as using machine learning methods to establish the relationship between pollen concentration and remote sensing data to predict future pollen concentrations [23–26]. However, neither the start date of flowering nor the concentration of pollen can fully reflect the outbreak date of pollen allergy, as they are only necessary conditions for pollen allergy rather than sufficient conditions. For example, the pollen concentration is also high in southern China in spring and autumn [27], but the concurrent number of pollen allergic patients is smaller than that in northern China [28]. This contradiction is because pollen allergy is caused by the allergenic pollen concentration rather than the total pollen concentration, and it is also related to seasonal changes in human immunity [29]. Therefore, it is necessary to establish a direct relationship between the incidence of pollen allergy (rather than flowering date or pollen concentration) and remote sensing data to predict the outbreak date of pollen allergy.

Beijing has various plants with allergenic pollen due to the warm temperate semi-humid climate [30], and the three main plants that cause pollen allergy in spring are cypress, poplar, and willow (especially cypress), whose flowering periods are all close to
each other [31], resulting in a high risk of pollen allergy [32]. Shi [33] found that one-quarter to one-third of respiratory allergy patients in Beijing are allergic to pollen. Wang et al. [34] discovered that the pollen concentration in Beijing has two peaks, spring (March to April) and summer-autumn (August to September), where allergic pollen mainly comes from woody plants in spring and herbaceous plants in summer-autumn. Meanwhile, the incidence of pollen allergies also has an obvious seasonal pattern, with a high incidence mainly in spring and autumn [35]. To benefit pollen allergic patients in Beijing, researchers have established a pollen concentration prediction model based on meteorological parameters since 2016 to predict recent pollen concentrations in all districts of Beijing [36]. However, the prediction model can only forecast pollen concentration at sites and tends to ignore the effects of other environmental factors on pollen concentration due to the limit of pollen monitor amounts and the uncomprehensive representativeness of meteorological parameters. To address this problem, some scholars have started to apply satellite data to build pollen concentration prediction models in Beijing. Bian et al. [23] established a next-day pollen concentration prediction model based on the average vegetation leaf area index (LAI) and daily meteorological data of tree and grass growth areas in Beijing with a nonlinear autoregressive neural network model (NARXnet). Although the prediction model has a high accuracy, the forecast period is very short, i.e., only for the next day but not for the medium or long term (e.g., the next 10 days or one month), and the forecast content is the total pollen concentration rather than the incidence of pollen allergy. Therefore, it is still difficult to aid the prevention of pollen allergies.

Although there are two peak periods for pollen allergy in Beijing, spring and summer-autumn, we focused on the spring pollen allergy and its vegetation phenological characteristics for prediction models, considering that the vegetation phenology changes more obviously in spring. This study aims to address two issues: first, to reveal the satellite-derived phenological characteristics of vegetation greenness on and before the outbreak date of spring pollen allergy in Beijing; and second, to establish a direct prediction model for the outbreak date of spring pollen allergy in Beijing based on the satellite-derived phenological characteristics of vegetation greenness.

2. Materials and Methods

2.1. Study Area

Beijing is surrounded by mountains in the west and north, and its middle is a plain open to the southeast (Figure 1). The plain in the urban area of Beijing with a similar vegetation phenology was selected as the study area (elevation ≤ 100 m) because the local climates and vegetation phenology differ significantly between the plain and the surrounding mountains in Beijing. The study area is located within 39°28′N–40°53′N, 115°83′E–117°30′E, with a warm-temperate semi-humid monsoon climate, an average annual temperature of 12 °C [37], an average annual precipitation of 621 mm [38], and diverse vegetation types. Wang et al. [39] investigated the pollen sources in Beijing urban area and found that the main pollen comes from buttercup (Ranunculaceae), amaranth (Amaranthaceae), pine (Pinaceae), cypress (Cupressaceae), elm (Ulmaceae), birch (Betula), artemisia (Artemisia), quinoa (Chenopodium), willow (Salix), humulus (Humulopsis), and Planetree (Platanus), and the highly allergenic pollen comes from Chinese pine (Pinus tabuliformis), cork oak (Quercus variabilis), green ash (Fraxinus pensylvanica), poplar (Populus tomentosa), tree of heaven (Ailanthus altissima), birch (Betula platyphylla), white elm (Ulmus pumila) and China savin (Juniperus chinensis). The types of allergenic pollen vegetation differ from season to season, with woody plants such as elm (Ulmaceae), cypress (Cupressaceae), pine (Pinaceae), and willow (Salix) in spring and herbaceous plants such as mulberry (Moraceae), chrysanthemum (Asteraceae), quinoa (Chenopodiaceae), and Graminae (Poaceae) in autumn [40].
2.2. Data

2.2.1. The Outbreak Dates of Spring Pollen Allergy

The outbreak dates of pollen allergies in Beijing were extracted from Sina Weibo records (see details in Appendix A). We obtained 13,404 pollen allergy Weibo records from 2011 to 2021 in Beijing in total, and 1803 valid records were manually screened for analysis. The outbreak dates of spring pollen allergies in Beijing were extracted from the maximum point on the second derivative of the fitted curve for Weibo data with the logistic function. The earliest date of the spring pollen allergy outbreak was on the 70th day (11th March), and the latest date was on the 91st day (1st April) in Beijing from 2011 to 2021. The extraction results were consistent with the outpatient records of pollen allergies in Beijing [34].

2.2.2. Remote Sensing Data

The satellite vegetation indices were downloaded from the Google Earth Engine (GEE). We selected the MODIS surface reflectance daily dataset (MOD09GA) with a resolution of 250 m from 2011 to 2021 to calculate different vegetation indices for different vegetation types during these 11 years.

2.2.3. Vegetation Classification Data

A vegetation classification map with a resolution of 500 m was produced based on satellite data in 2020 and 2021 and the FROM-GLC10 land cover map released by Tsinghua University, which classified the study area into evergreen forest, deciduous forest, grassland, cropland, and non-vegetation areas (Figure 1). The details of specific vegetation classification data and methods are listed in Appendix B. This study focuses on two vegetation types, evergreen forest and deciduous forest, which are related to spring pollen allergens.
2.3. Methods
2.3.1. Extraction of Satellite-Derived Phenological Characteristics of Vegetation Greenness within the 30 Days before the Spring Average Pollen Allergy Outbreak Date during 2011–2021

The concentration and type of pollen are related to the vegetation type in a region [41]. Given that the allergenic vegetation in Beijing is mainly woody plants in spring [42], we selected the vegetation indices in forests to calculate the greenness characteristics before the outbreak date of pollen allergy in spring.

With reference to the study results of pollen identification based on remote sensing spectral features by Peng et al. [12], we used the normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), 2-band enhanced vegetation index (EVI2), the sum of blue, red, and near-infrared reflectance (NIR + R + B) and the ratio of green to red reflectance (G/R) to analyze the characteristic.

For a certain vegetation type, the satellite-derived phenological characteristic of vegetation greenness is defined as a combination of a certain vegetation index value and its corresponding date. The technical process for the extraction of the greenness characteristics before pollen allergy outbreak is shown in Figure 2, which includes the following three main steps: (1) The preprocessing of remote sensing vegetation index time-series data; (2) the extraction of satellite-derived phenological characteristics of vegetation greenness; and (3) the analysis of the representative capacity of the greenness characteristics to the outbreak date of pollen allergy.

(1) Preprocessing of remote sensing vegetation index time-series data

The daily remote sensing images from 2011 to 2021 were used to calculate the vegetation index value for each vegetation type. Since there are many noises in the daily time-series data, we adopted two methods to reduce the noises. First, the spatial 5% trimmed mean of each vegetation index value in every vegetation type was taken as the daily phenological characteristics for each year. Second, the data were smoothed by taking the mean of the higher 6 values within the previous 11-day filtering window for each day. Since we have set the forecast of pollen allergies to be 30 days in advance and the earliest date of spring pollen allergies in Beijing from 2011 to 2021 is the 70th day and the latest date is the 91st day, we only need to consider the vegetation index time-series data from the 50th to 90th day of each year (Figure 2c). Based on the shape of the vegetation index time-series curve from the 50th to the 90th day, as shown in Figure 2c, we chose a linear function to fit the daily time-series curve and the cumulative time-series curve of the vegetation index.

\[ Y(t) = m + nt \]  

where \( Y(t) \) is the daily vegetation index value or cumulative vegetation index value; \( t \) is DOY 50 \( \leq t \leq 90; \) and \( m, n \) are the parameters that need to be fitted.

(2) The extraction of satellite-derived phenological characteristics of vegetation greenness

The outbreak dates of spring pollen allergy extracted from the Weibo data were used to find the corresponding vegetation index value on the linear fitted curve. For each vegetation type, the value of a certain vegetation index on the outbreak date of pollen allergy is regarded as the satellite-derived phenological characteristic of vegetation greenness in this study (right plots in Figure 2c).

(3) Analysis of the representative capacity of the greenness characteristics to the outbreak date of pollen allergy

To determine the greenness characteristic that can best reflect spring pollen allergy outbreaks, we calculated the sensitivity coefficients of each characteristic to the outbreak date of spring pollen allergies and the correlation coefficients between each characteristic and the outbreak date of spring pollen allergies.
Figure 2. Technical process diagram for the extraction of satellite-derived vegetation greenness phenological characteristics before the pollen allergy outbreak. T in the right plots of (c) is the average DOY of spring pollen allergy outbreak during 2011–2021; VI in the right plots of (c) and (d) is the vegetation index value corresponding to the average DOY of pollen allergy outbreak; x₁ in the left plot of (c) is the DOY that 20 days earlier than the earliest breakout date of spring pollen allergy during 2011–2021, i.e., the 50th day; x₂ is the average DOY of spring pollen allergy outbreak during 2011–2021, i.e., the 81st day; y₁ is the vegetation index value corresponding to x₁ on the fitted curve of the smoothed multiyear average daily curve of a vegetation index for a vegetation type; y₂ is the vegetation index value corresponding to x₂ on the fitted curve of the smoothed multiyear average daily curve of a vegetation index for a vegetation type; tₙ in the right plot of (d) is the corresponding DOY of VI in each year during 2011–2021; and Tₙ is the DOY of spring pollen allergy outbreak in each year during 2011–2021.
The sensitivity coefficient can reflect the sensitivity of each satellite-derived phenological characteristic of vegetation greenness to the outbreak date of spring pollen allergy. It is calculated as follows:

\[
S_V = \frac{(y_2 - y_1)/y_1}{(x_2 - x_1)/x_1}
\]

(2)

where \(x_1\) is the DOY 20 days earlier than the earliest breakout date of spring pollen allergy during 2011–2021, i.e., the 50th day; \(x_2\) is the average DOY of spring pollen allergy outbreak during 2011–2021, i.e., the 81st day; \(y_1\) is the vegetation index value corresponding to \(x_1\) on the fitted curve of the smoothed multiyear average daily curve of a vegetation index for a vegetation type; and \(y_2\) is the vegetation index value corresponding to \(x_2\) on the fitted curve of the smoothed multiyear average daily curve of a vegetation index for a vegetation type (the right plot in Figure 2d).

The correlation coefficient can reflect the degree of consistency between the interannual variation of each characteristic and the outbreak date of spring pollen allergy. A higher value of the correlation coefficient indicates a more stable ability to represent the phenological characteristics of spring pollen allergies. To calculate the correlation coefficient, the average outbreak date (\(T\)) of spring pollen allergy (\(T_{2011}, T_{2012}, \ldots, T_{2021}\)) during 2011–2021 was used to find the corresponding DOY (\(t_{2011}, t_{2012}, \ldots, t_{2021}\)) in each vegetation index value time series of different vegetation types every year. Then, the Pearson correlation coefficient between the real outbreak date (\(T_{2011}, T_{2012}, \ldots, T_{2021}\)) and the corresponding satellite DOY (\(t_{2011}, t_{2012}, \ldots, t_{2021}\)) was calculated for each combination of vegetation indices and vegetation types (the left plot in Figure 2d).

2.3.2. Establishment and Accuracy Assessment of the Prediction Models

We established prediction models based on the satellite-derived phenological characteristics of vegetation greenness within the 20 days before the earliest date of the outbreak dates of spring pollen allergy in Beijing during 2011–2021 (Figure 3a,b). The prediction models were developed as follows.

\[
Y = mt + n
\]

\[
W = \Delta t = t - t_0 = (Y - n)/m - 81
\]

(3)

(4)

where \(Y\) is the preprocessed multiyear average daily fit curve of the remote sensing vegetation index when establishing the prediction model or the preprocessed daily vegetation index in a given year when predicting; \(m\) and \(n\) are the coefficient of the linear fit of \(Y\); \(W\) is the prediction function, \(\Delta t\) is the predicted number of days to the outbreak date of spring pollen allergy in Beijing (a negative value indicates that the outbreak date has not yet arrived, and a positive value indicates that the outbreak date has passed); \(t\) is DOY of the current date; and \(t_0\) is DOY of the average outbreak date of spring pollen allergy in Beijing during 2011–2021, i.e., the 81st day. The prediction models for the outbreak date of spring pollen allergy were named the linear fit prediction model and cumulative linear fit prediction model, respectively, according to the different preprocessing of remote sensing vegetation index data.

The accuracy of the prediction models was evaluated with the root mean square error (RMSE), which reflects the deviation of the forecast date from the real date. A smaller RMSE means a higher forecast accuracy. To test the performance of each prediction model (the linear fit prediction model and cumulative linear fit prediction model), the vegetation index on the 50th day of each year during 2011–2021 (the DOY of the 20 days before the earliest outbreak date of spring pollen allergy in these 11 years) was put into every prediction model to calculate the countdown to the outbreak date of spring pollen allergy in Beijing, and then the predicting outbreak date of pollen allergy for each year in Beijing was determined (Figure 3c,d). Finally, the RMSE between forecasted dates and real dates was calculated to select the best prediction model.
To give an objective and fair evaluation of the models, we established the models by randomly selecting 7 years from 2011–2021 and predicted outbreak dates of pollen allergy for the remaining 4 years. A total of 330 training sessions (C_11^7 = 330) were performed for each model (the linear fit prediction model and cumulative linear fit prediction model).

3. Results

3.1. Satellite-Derived Phenological Characteristics of Vegetation Greenness within 30 Days before the Spring Average Pollen Allergy Outbreak Date during 2011–2021

The value range of each greenness characteristic derived from different vegetation indices was diverse in each vegetation type within the 30 days before the pollen allergy outbreak date (Table 1). During spring, the NIR + R + B values of each vegetation type had the most obvious magnitude of variation among all indices during the 30 days before the outbreak date, and the most obvious change in its value occurred in deciduous forest, from 1.800 to 2.061. The G/R values of each vegetation type changed the least, among which the G/R value of evergreen + deciduous forest changed by only 0.006 during the 30 days.
Table 1. The value range of each greenness characteristic derived from different vegetation indices within the 30 days before the spring average pollen allergy outbreak date during 2011–2021.

<table>
<thead>
<tr>
<th>Vegetation Type</th>
<th>NDVI</th>
<th>EVI</th>
<th>EVI2</th>
<th>NIR + R + B</th>
<th>G/R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evergreen Forest</td>
<td>0.257–0.309</td>
<td>0.125–0.131</td>
<td>0.113–0.138</td>
<td>2.074–2.217</td>
<td>0.967–0.974</td>
</tr>
<tr>
<td>Deciduous Forest</td>
<td>0.196–0.227</td>
<td>0.107–0.123</td>
<td>0.095–0.126</td>
<td>1.800–2.061</td>
<td>0.971–0.976</td>
</tr>
<tr>
<td>Evergreen + Deciduous Forest</td>
<td>0.233–0.260</td>
<td>0.117–0.135</td>
<td>0.113–0.124</td>
<td>2.008–2.139</td>
<td>0.969–0.975</td>
</tr>
</tbody>
</table>

In terms of sensitivity coefficients (Table 2), the sensitivity coefficient of EVI2 to the outbreak date of spring pollen allergy was the largest at 0.270 in evergreen forest, followed by NDVI of evergreen forest with a sensitivity coefficient of 0.253, and the sensitivity coefficients of G/R of all vegetation types were smaller than 0.01.

Table 2. The sensitivity coefficients of each greenness characteristic to the vegetation greenness date of spring pollen allergy.

<table>
<thead>
<tr>
<th>Vegetation Type</th>
<th>NDVI</th>
<th>EVI</th>
<th>EVI2</th>
<th>NIR + R + B</th>
<th>G/R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evergreen Forest</td>
<td>0.253</td>
<td>0.060</td>
<td>0.270</td>
<td>0.086</td>
<td>0.009</td>
</tr>
<tr>
<td>Deciduous Forest</td>
<td>0.198</td>
<td>0.187</td>
<td>0.116</td>
<td>0.181</td>
<td>0.006</td>
</tr>
<tr>
<td>Evergreen+ Deciduous Forest</td>
<td>0.145</td>
<td>0.192</td>
<td>0.122</td>
<td>0.082</td>
<td>0.008</td>
</tr>
</tbody>
</table>

Note: Bold font indicates the maximum value of the sensitivity coefficient in each row.

The correlation coefficients between each greenness characteristic and the outbreak date of pollen allergy are shown in Table 3. The NDVI and EVI2 of each vegetation type were significantly correlated \( (p < 0.05) \) with the outbreak date of spring pollen allergy, with the largest correlation coefficient of 0.762 for the NDVI in deciduous forest, while the EVI and NIR + R + B of each vegetation type were not significantly correlated with the outbreak date.

Table 3. The correlation coefficients between the interannual variations in each greenness characteristic and the outbreak dates of pollen allergy.

<table>
<thead>
<tr>
<th>Vegetation Type</th>
<th>NDVI</th>
<th>EVI</th>
<th>EVI2</th>
<th>NIR + R + B</th>
<th>G/R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evergreen Forest</td>
<td>0.724 *</td>
<td>0.289</td>
<td>0.693 *</td>
<td>0.159</td>
<td>0.612 *</td>
</tr>
<tr>
<td>Deciduous Forest</td>
<td>0.762 *</td>
<td>0.429</td>
<td>0.633 *</td>
<td>0.485</td>
<td>0.538</td>
</tr>
<tr>
<td>Evergreen+ Deciduous Forest</td>
<td>0.717 *</td>
<td>0.675</td>
<td>0.705 *</td>
<td>0.280</td>
<td>0.519</td>
</tr>
</tbody>
</table>

Note: Bold font indicates the maximum value of the correlation coefficient in each row; * indicates significance at \( p = 0.05 \).

Combining the sensitivity coefficients and correlation coefficients of each greenness characteristic, we found that the date of EVI2 in evergreen forest first reaching 0.137 can best reflect the outbreak date of pollen allergy in spring.

3.2. The Prediction Models and Their Accuracies

Based on the EVI2 of evergreen forest during 2011–2021, the linear fit prediction model and cumulative linear fit prediction model were developed, which best reflects the outbreak date of spring pollen allergy in Beijing (Figure 4). The expressions of each prediction model were obtained as follows:
Combining the sensitivity coefficients and correlation coefficients of each greenness characteristic, we found that the date of EVI2 in evergreen forest first reaching 0.137 can best reflect the outbreak date of pollen allergy in spring.

3.2. The Prediction Models and Their Accuracies

Based on the EVI2 of evergreen forest during 2011–2021, the linear fit prediction model and cumulative linear fit prediction model were developed, which best reflects the outbreak date of spring pollen allergy in Beijing (Figure 4). The expressions of each prediction model were obtained as follows:

(1) The linear fit prediction model

\[ W = 1666.67Y_1 - 219.5 \]  \( \text{(5)} \)

where \( Y_1 \) is the smoothed daily EVI2 value of evergreen forest for a given forecast year and \( W \) is the number of days to the outbreak date of spring pollen allergy in Beijing.

(2) The cumulative linear fit prediction model

\[ W = 8.628Y_2 - 78.69 \]  \( \text{(6)} \)

where \( Y_2 \) is the cumulative smoothed daily EVI2 value of evergreen forest for a given forecast year and \( W \) is the number of days to the outbreak date of spring pollen allergy in Beijing.

The prediction accuracies obtained from 330 training sessions for each model based on EVI2 of evergreen forest are shown in Table 4. The average RMSEs of the linear fit prediction model and the cumulative linear fit prediction model were 95.549 days and 3.589 days, respectively, which indicated that the linear fit prediction model has little predictive power.
while the cumulative linear fit prediction model has a very good prediction ability. It should be noted that the first training session in Table 4 with the years 2011–2017 for model building and the years 2018–2021 for model test can also get a low RMSE (i.e., 3.369 days) for the cumulative linear fit prediction model, though there is a dramatic increase in the number of valid pollen allergy Weibo records since 2018 (Figure A1), which indicated that the dramatic increase in the number of valid records since 2018 did not affect the prediction accuracy.

Table 4. Cross-test results of the prediction models based on EVI2 of evergreen forest, NDVI of evergreen forest and EVI2 of deciduous forest for 330 training sessions.

<table>
<thead>
<tr>
<th>Years for Model Building</th>
<th>Years for Model Test</th>
<th>RMSE for the Linear Fit Prediction Model</th>
<th>RMSE for the Cumulative Linear Fit Prediction Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>EVI2 of Evergreen Forest</td>
<td>NDVI of Evergreen Forest</td>
</tr>
<tr>
<td>Average of RMSEs (mean ± sd)</td>
<td></td>
<td>95.549 ± 197.572</td>
<td>92.636 ± 134.075</td>
</tr>
</tbody>
</table>

To further verify the validity of the screened vegetation phenological greenness characteristic that best reflects the date of pollen allergy outbreak (i.e., EVI2 of evergreen forest), we also tested the prediction accuracies for the models based on NDVI of evergreen forest and EVI2 of deciduous forest (Table 4), respectively. Controlling the variable in model construction (i.e., change EVI2 to NDVI or change evergreen forest to deciduous forest), we found that the prediction accuracy of each model built by the above two vegetation characteristics was lower than that of the prediction model built by EVI2 of evergreen forest.

4. Discussion

4.1. Phenological Characteristics of Remote Sensing Vegetation Greenness at the Beginning and Early Stages of the Spring Pollen Allergy Outbreak in Beijing

We found that the date of EVI2 in evergreen forest first reaching 0.137 can best reflect the outbreak date of pollen allergy in spring. Evergreen forest can better reflect spring pollen allergy outbreaks in Beijing than other vegetation types because cypress, the main allergenic vegetation in Beijing in spring, is an evergreen plant [31]. EVI2 can better indicate vegetation greenness before flowering because it can reduce atmospheric and soil disturbances [43,44], making it more sensitive to vegetation greenness in areas with high background noises [44,45].

EVI2 at 0.137 of evergreen forest has a specific phenological indication. The remote sensing vegetation index time-series curve for a complete vegetation growth season generally has four key transition dates: greenup (the date of onset of photosynthetic activity), maturity (the date at which plant green leaf area achieves maximum), senescence (the date at which photosynthetic activity and green leaf area begin to rapidly decrease), and dormancy (the date at which physiological activity becomes near zero) [6]. In this study, the satellite-derived greenness characteristic that best reflected the outbreak date of spring pollen allergy corresponded to the key greenup transition date: the date of EVI2 at 0.137 in the growth period of evergreen forest in spring corresponded to the greenup date of evergreen forest (e.g., the date of EVI2 at 0.136 of evergreen) (Figure 5). Actually, the allergic pollens in Beijing in spring mainly come from cypress, poplar, and willow, especially cypress [40]. As an evergreen forest, cypress flowerings at the time of needle-leaf flush, while poplar and willow flowering first and then leaf out immediately. However, if the flowering and leaf spreading periods of the allergic plant are separated by a long time, its greenness characteristics cannot reflect its flowering, and the method will be not applicable.
4.2. Advantages of the Prediction Models

The cumulative linear fit prediction model has a very good prediction ability for the outbreak date of spring pollen allergy in Beijing based on the EVI2 of evergreen forest. It not only attenuates the fluctuation of the original data before the pollen allergy outbreak by cumulating daily data, but also has a higher fitting accuracy compared with the linear fit prediction model, in which the cumulative linear fitting curve almost coincides with the cumulative EVI2 time-series curve at the prediction period (Figure 4).

The existing pollen allergy-related prediction models mainly forecast pollen allergy outbreaks by forecasting the airborne pollen concentration through the empirical relationships between pollen concentrations and meteorological elements [46,47]. There are also models utilizing the flowering period of allergenic vegetation estimated by the blooming period of earlier flowering vegetation [48] to predict pollen allergies, but these models only predict vegetation characteristics related to pollen allergies and do not directly predict the date when people will be affected by allergic pollens. In addition, the existing prediction models require a substitution of data related to multiple meteorological factors for actual forecasting, which greatly limits the application of the prediction models. Moreover, the forecasting accuracy of the existing models is reduced by this indirect relationship between forecasting features and pollen allergy outbreaks.

The newly developed cumulative linear fit prediction model is not only capable of directly forecasting pollen allergy outbreaks, but also has a high data availability to operate, which only requires remote sensing vegetation index time-series data in the preliminary period of pollen allergy outbreaks. Moreover, the cumulative linear fit prediction model is more representative (directly reflecting pollen allergy rather than pollen concentration) and operational, and the accuracy of this direct forecast is higher (RMSE of 3.5 days) than that of the indirect forecast.

4.3. The Importance of Data Preprocessing for the Daily Vegetation Index Time-Series Data and Limitations for the Application of the Prediction Models

There are many noises in the daily vegetation index time-series data because of the disturbances from cloud contamination and atmospheric conditions [49]. Therefore, it is needed to reduce the noises before the application of the prediction models. In this study, the spatial 5% trimmed mean of each vegetation index value in every vegetation type was
taken as the daily phenological characteristics for each year, and then the data were further smoothed by taking the mean of the higher six values within the previous 11-day filtering window for each day. Other noise reduction methods may affect the forecast accuracy since the prediction models are easily limited to the quality of remote sensing data. In extreme cases, if the study area is long affected by cloud cover, which causes degradation in the quality of remote sensing data, the forecast accuracy will be significantly reduced.

5. Conclusions

This study revealed the satellite-derived phenological characteristics of vegetation greenness before the spring pollen allergy outbreak in Beijing based on 11-year Sina Weibo data and corresponding satellite data, and established a prediction model to forecast the outbreak date of spring pollen allergies in Beijing based on the phenological characteristics that best reflect spring pollen allergies. The main conclusions are as follows.

(1) The satellite-derived phenological characteristics of vegetation greenness are obvious during the early period of the spring pollen allergy outbreak in Beijing. The date of EVI2 in evergreen forest first reaching 0.138 can best reflect the outbreak date of pollen allergy in spring. Moreover, it has a specific phenological indication: the date of EVI2 in evergreen forest first reaching 0.138 in spring basically corresponds to the greenup date.

(2) The cumulative linear fit prediction model based on EVI2 of evergreen forest has a very good prediction ability for forecasting the spring pollen allergy outbreak date in Beijing, and it can predict 30 days in advance, with a low average RMSE of 3.6 days. The existing forecast models of pollen allergy outbreak are mainly based on indirect forecasting of pollen concentrations, while the newly developed model can forecast pollen allergy outbreak directly. It only requires the time-series data of remote sensing vegetation index before pollen allergy outbreak to achieve the forecast, which is more representative (directly reflecting pollen allergy rather than pollen concentration) and operable. However, the forecast accuracy of the cumulative linear fit prediction model is easily limited by the quality of remote sensing data, and if the study area is affected by cloud cover for a long time, which results in degradation of the quality of satellite remote sensing data, the forecast accuracy will be significantly reduced.

Author Contributions: W.Z. and X.Y. were responsible for the conceptualization of the study and analysis of the data; C.Z. is contributed in writing and reviewing. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by Graduate-Postgraduate Connection Talent Cultivation Program of the Faculty of Geographical Science (FGS), Beijing Normal University (BNU) and the National Natural Science Foundation of China Major Program (No. 42192580, 42192581).

Data Availability Statement: Data is available upon request from the corresponding author.

Acknowledgments: Thanks for the help provided by the whole group of professor Zhu W., especially He B.

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

Appendix A. Extraction of the Start Dates of Pollen Allergy Outbreaks in Beijing Based on Sina Weibo Data

Appendix A.1. Data

A total of 13,404 Weibo records were retrieved by a self-written Python program with the keyword “pollen allergy” in Beijing from 2011–2021 from the Sina Weibo website (https://weibo.com/ (accessed on 10 September 2021)). After visual interpretation of the retrieved Weibo contents, 11,601 invalid Weibo records (e.g., microblogs published in Beijing and without pollen allergy symptoms at the time of publication, such as advertisements) were excluded, and 1803 valid Weibo records (i.e., microblogs published in Beijing and with pollen allergy symptoms at the time of publication) were screened for the extraction of pollen allergy outbreak dates. The statistical results showed that the number of valid
records varies little from 2011–2017 (36–99) but has increased dramatically since 2018 (299–395) (Figure A1), which is attributed to the wide popularity of Weibo on mobile phones since 2018. Since the pollen allergy outbreak dates were extracted separately based on the Weibo data of each year, the dramatic increase in the number of valid records since 2018 does not affect the overall extraction results.

![Figure A1](image-url)

**Figure A1.** Weibo data with the keyword “pollen allergy” in Beijing during 2011–2021. The sharp increase in Weibo data in 2018 is because the mobile side of Weibo began to be widely popular in that year, with mobile active users accounting for 93% of the total active users, which is sourced from the Weibo Data Center. Weibo 2018 User Development Report [EB/OL], 2019-03-15.

**Appendix A.2. Methods**

The number of valid records per 5 days from 2011 to 2021 was counted, and the dates of valid records were recorded in DOY. The data were fitted with a logistic curve (Figure A2). In the process of fitting, each peak of the curve was divided into two parts, and the left part was fitted. The best fitting curve was selected according to the principle of minimum fitting variance. Finally, the local maximum point of the second derivative of the best fitting curve in the left part was used as the outbreak date of pollen allergy in that year.

![Figure A2](image-url)

**Figure A2.** A schematic diagram for extracting the outbreak dates of pollen allergies based on valid microblog data.
Appendix A.3. Results

The outbreak dates of spring pollen allergy in Beijing were from March 11th to April 1st (Table A1), and the peak of spring pollen allergy diagnosis in Beijing also occurred from March to April (Table A2), which indicated that the outbreak dates of spring pollen allergy in Beijing extracted from the Weibo data were reliable and could be used to establish the prediction models. However, since the diagnosis data is on a monthly scale, we can only use this data to compare with the pollen allergy outbreak dates extracted from the Weibo on a monthly scale (the peaks are all in March), and cannot verify the consistency of interannual variation at a shorter time scale (e.g., 5 or 10 days).

Table A1. The start date of spring pollen allergy outbreak extracted from microblog data.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Day of year</td>
<td>86</td>
<td>87</td>
<td>91</td>
<td>81</td>
<td>82</td>
<td>76</td>
<td>81</td>
<td>81</td>
<td>70</td>
<td>72</td>
<td>88</td>
</tr>
<tr>
<td>Date</td>
<td>March</td>
<td>March</td>
<td>April</td>
<td>March</td>
<td>March</td>
<td>March</td>
<td>March</td>
<td>March</td>
<td>March</td>
<td>March</td>
<td>March</td>
</tr>
</tbody>
</table>

Note: The peak of spring pollen allergy diagnosis in Beijing occurred from March to April.

Table A2. Diagnosis data of pollen allergies in spring in Beijing in 2015 [34].

<table>
<thead>
<tr>
<th>Month</th>
<th>Allergic Rhinitis</th>
<th>Bronchial Asthma</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan.</td>
<td>1429</td>
<td>1216</td>
<td>2645</td>
</tr>
<tr>
<td>Feb.</td>
<td>1285</td>
<td>963</td>
<td>2248</td>
</tr>
<tr>
<td>Mar.</td>
<td>2554</td>
<td>1263</td>
<td>3817</td>
</tr>
<tr>
<td>Apr.</td>
<td>2021</td>
<td>1279</td>
<td>3300</td>
</tr>
<tr>
<td>May</td>
<td>1728</td>
<td>1161</td>
<td>2889</td>
</tr>
<tr>
<td>June</td>
<td>1785</td>
<td>1232</td>
<td>3017</td>
</tr>
<tr>
<td>July</td>
<td>1451</td>
<td>1165</td>
<td>2616</td>
</tr>
<tr>
<td>Aug.</td>
<td>3343</td>
<td>1471</td>
<td>4814</td>
</tr>
<tr>
<td>Sept.</td>
<td>2744</td>
<td>1465</td>
<td>4209</td>
</tr>
<tr>
<td>Oct.</td>
<td>2005</td>
<td>1238</td>
<td>3243</td>
</tr>
<tr>
<td>Nov.</td>
<td>1849</td>
<td>1288</td>
<td>3137</td>
</tr>
<tr>
<td>Dec.</td>
<td>2300</td>
<td>1513</td>
<td>3813</td>
</tr>
</tbody>
</table>

Note: The peak of spring pollen allergy diagnosis in Beijing occurred from March to April.

Appendix B. Description of Vegetation Classification

Appendix B.1. Data

Sentinel-2 images, the vegetation index product of MODIS (MOD13A1), and the Fromglc10_2017v01 land cover classification product were used for vegetation classification. Sentinel-2 data were obtained from the ESA website (https://scihub.copernicus.eu/ (accessed on 10 January 2022)) with a spatial resolution of 10 m. The Sentinel-2 images in August 2020 were used to identify grassland in the study area, and the images in November 2021 were used to identify evergreen forest in the study area. The MOD13A1 data were obtained from the USGS website (https://lpdaacsvc.cr.usgs.gov/ (accessed on 20 January 2022)), with a spatial resolution of 500 m. Its NDVI data in July 2021 were used to identify vegetation cover areas in the study area. The Fromglc10_2017v01 land cover classification product comes from Tsinghua University (http://data.ess.tsinghua.edu.cn/fromglc10_2017v01.html (accessed on 5 March 2022)), which was produced based on the 2017 Sentinel-2 data with a spatial resolution of 10 m. The cropland class from these data was directly used in this study.

The training and testing samples used for vegetation classification were selected by visual interpretation of high-resolution Google Earth images. For the classification of vegetation and non-vegetation, 230 training samples were selected (180 vegetation samples and 50 non-vegetation samples), and 220 testing samples were selected (175 vegetation samples and 45 non-vegetation samples); for the classification of grassland and forest, 55 training samples were selected (25 grassland samples and 30 forest samples), and
60 testing samples were selected (30 grassland samples and 30 forest samples); and for the classification of evergreen and deciduous forests, 120 training samples were selected (50 for evergreen forests and 70 for deciduous forests) and 130 testing samples were selected (55 evergreen forest samples and 30 forest samples).

Appendix B.2. Methods

The vegetation classification process is shown in Figure A3, and the general idea is to extract them by type.

![Process diagram of vegetation classification.](image)

When classifying the vegetation and non-vegetation types, the NDVI threshold method was used because of the large differences in NDVI values between vegetation and non-vegetation in summer. The threshold value was determined by the cumulative frequency probability distribution plot method based on the NDVI data in July 2021, and the NDVI value corresponding to the cumulative frequency probability of 90% was taken as the threshold value (e.g., the value of NDVI at 0.37) (Figure A4). The purpose of vegetation
classification in this study is to analyze the vegetation phenological characteristics of pollen allergies, so it is necessary to pursue a high user accuracy of vegetation classification, and using this method to determine the classification threshold can theoretically ensure that the user accuracy of vegetation classification is higher than 90%. Although the producer accuracy of vegetation classification may be reduced (i.e., some pixels that are actually vegetation are not classified as vegetation), this does not affect the subsequent analysis of vegetation phenological characteristics of pollen allergies.

When classifying cropland, grassland, and forest, cropland was first masked from the Fromglc10_2017v01 land cover data. Since the reflectance of forest and grassland differed greatly in the red band, the near-infrared band and three red edge bands in summer, forest, and grassland were identified based on these bands of Sentinel-2 data in August by random forest classification.

When classifying the evergreen and deciduous forests, the NDVI threshold method was used because the NDVI values of evergreen and deciduous forests differed greatly in autumn and winter. The specific procedure was the same as that used to classify vegetation and non-vegetation, and the frequency probability distribution of NDVI values for the training samples of evergreen and deciduous forests is shown in Figure A5. The final NDVI threshold value for classifying evergreen and deciduous forests was determined to be 0.67.

Finally, the classification accuracy is obtained by calculating the confusion matrix with testing samples.
Figure A4. Pixel frequency probability distribution of NDVI values of vegetation and non-vegetation training samples.

When classifying cropland, grassland, and forest, cropland was first masked from the Fromglc10_2017v01 land cover data. Since the reflectance of forest and grassland differed greatly in the red band, the near-infrared band and three red edge bands in summer, forest, and grassland were identified based on these bands of Sentinel-2 data in August by random forest classification. When classifying the evergreen and deciduous forests, the NDVI threshold method was used because the NDVI values of evergreen and deciduous forests differed greatly in autumn and winter. The specific procedure was the same as that used to classify vegetation and non-vegetation, and the frequency probability distribution of NDVI values for the training samples of evergreen and deciduous forests is shown in Figure A5. The final NDVI threshold value for classifying evergreen and deciduous forests was determined to be 0.67.

Figure A5. Pixel frequency probability distribution of NDVI values of evergreen and deciduous forest training samples.

Appendix B.3. Results

The results of vegetation classification in the study area are shown in Figure 1, and the user accuracy and producer accuracy of each vegetation type are shown in Table A3. The user accuracy is higher than 90%, which means that the classified vegetation types are reliable; and the producer accuracy is lower than the user accuracy, but it is above 80%. Since this study mainly uses the vegetation classification results to analyze the satellite phenological characteristics of vegetation greenness before pollen allergy outbreaks, even if some vegetation types are not identified, it does not affect the research results, so a user accuracy above 90% is sufficient to meet the needs of this study.

Table A3. The accuracy of vegetation classification.

<table>
<thead>
<tr>
<th>Vegetation Type</th>
<th>User Accuracy/%</th>
<th>Producer Accuracy/%</th>
<th>Number of Pixels</th>
<th>Area/km²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-vegetation area</td>
<td>99.83</td>
<td>92.47</td>
<td>13408</td>
<td>3352.0</td>
</tr>
<tr>
<td>Grassland</td>
<td>95.86</td>
<td>84.79</td>
<td>2790</td>
<td>697.5</td>
</tr>
<tr>
<td>Evergreen forest</td>
<td>97.53</td>
<td>94.87</td>
<td>367</td>
<td>91.8</td>
</tr>
<tr>
<td>Deciduous forest</td>
<td>91.34</td>
<td>82.93</td>
<td>607</td>
<td>151.8</td>
</tr>
<tr>
<td>Cropland</td>
<td>90.76</td>
<td>85.65</td>
<td>9644</td>
<td>2411.0</td>
</tr>
</tbody>
</table>

Note: The size of each pixel is 500 m x 500 m.

References


