# Research on Predicting Spread of Dandelion Seed Based on PCA and Dandelion Optimizer

Qihao Chen *, Yuxuan Zhang, Ruochong Zhang and Jiaxuan Zheng

The Affiliated High School of Peking University, Beijing, China

* Corresponding Author Email: chenqihao2025@i.pkuschool.edu.cn

**Abstract.** Plants such as Taraxacum officinale, as known as dandelion, flood our life and therefore influence our world drastically. They can result in both beneficial outcomes and negative consequences. As a major component of invasive species, the spread of dandelions in the ecosystem could contribute to the broke of ecological balance, causing economic loss in agriculture. Therefore, researching the spread and influencing factors of dandelions has extremely significant importance. To determine the impact factors, we comprehensively consider complicated sub-factors. After quantifying and unifying the data, we utilize PCA for low-dimensional analysis. And, we introduce data from other invasive species to validate the generalization of our model. Based on the results of PCA, we establish a model based on Dispersion Optimization (DO), which predicts the spread of dandelions under different climatic conditions for 1 month, 2 months, 3 months, 6 months, and 12 months by collecting statistical data on climate indicators. We found that the longer the time period, the smaller the differences in dandelion spread across different climates.

**Keywords:** Taraxacum officinale, seed distribution, Dandelion Optimizer (DO), Principal Component Analysis (PCA).

## 1. Introduction

Taraxacum officinale, also known as the dandelion, is considered as a friend or a foe for human beings and the whole ecosystem. Compared to the spread of other plants, the spread of dandelions in its “puffball” stage is quite random since it could mainly be influenced by the wind. Furthermore, the relationship between dandelions, humans, and other flora is complex because dandelions could both influence the ecosystem negatively and positively [1].

In this paper, we focus on studying the spread of dandelions and the relationship between invasive species, particularly dandelions, and ecosystems. We design a mathematical model based on PCA [2] to assess the impact factors of most invasive species. Through the analysis of impact factors, we build a model based on dandelion optimizer [3] to predict the spread range of dandelions, thus providing scientific basis for managing and controlling the spread of dandelions. Additionally, our research can help people better understand the complex relationship between invasive species and ecosystems, providing reference and guidance for environmental protection efforts.

## 2. Analysis of Impact Factor

### 2.1. Subfactor

Before taking advantage of our model, we have to determine the specific function of every factor we consider. The following matrix contains all five subfactors we consider:

$$[Z_c \ \ Z_n \ \ Z_d \ \ Z_v \ \ Z_s] \tag{1}$$

Here, $Z_c$ quantifies the species' adaptability and competitive success within the environental carrying capacity, $Z_n$ represents the fecundity through the number of seeds produced, $Z_d$ measures the dispersal range reflecting the potential area of impact, $Z_v$ assesses the root volume as an indicator of resource competition, and $Z_s$ evaluates the species' utility value to humans [4].

The factors $Z_n$ and $Z_d$ can be mathematically expressed with the following equations:

$$Z_n = (np\%)^{\frac{365}{T}} \tag{2}$$

$$Z_d = e\left(\frac{365}{T} r_{max}\right)^2 \tag{3}$$

**2.2. Principal Component Analysis (PCA)**

Principal Component Analysis (PCA) is a multivariate analysis to evaluate the key, or principal, component in our impact factor computation. In our model, we use it to take advantage of the method to only utilize a small number of variables to describe a deviation of the whole database (principal component).

To begin with, we utilize Equation 4 to calculate the covariance of factors: We can use it to measure the relationship between two factors in our set. Besides, if the covariance is positive, both variables increase or decrease simultaneously; In contrast, if the covariance is negative, the change direction of the two factors is different.

$$Cov(Z) = \frac{1}{n-1}(Z-\bar{Z})^T(Z-\bar{Z}) \tag{4}$$

In the equation above, $\bar{Z}$ is the mean vector for each column of the matrix Z. Then, we caculate the covariance matrix of all the factors we coonsider.

$$\Sigma = \begin{bmatrix} Cov(z_c, z_c) & Cov(z_c, z_n) & \cdots & Cov(z_c, z_s) \\ Cov(z_n, z_c) & Cov(z_n, z_n) & \cdots & Cov(z_n, z_s) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(z_s, z_c) & Cov(z_s, z_n) & \cdots & Cov(z_s, z_s) \end{bmatrix} \tag{5}$$

In the equation above, Cov $(Z_i, Z_j)$ is the covariance of factors $z_i$ and $z_j$

Then, we need to take advantage of the eigenvalue to separate the covariance matrix above.

$$Z = VDV^{-1} \tag{6}$$

For the Equation 6, V is a matrix containing all the eigenvectors of Z, and D is a diagonal matrix whose elements on the diagonal are the corresponding eigenvalues.

After the computation, we attain the principal components of our model: environmental capacity, number of seeds, and root volume, corresponding the contribution rate 61.12%, 20.43%. To specifically calculate the impact factor for invasive species, we get the principal component function for each of the principal component mentioned above.

$$F_1 = 0.6613Z_1 + 0.5834Z_2 - 0.0450Z_3 - 0.3967Z_4 - 0.2509Z_5 \tag{7}$$

$$F_2 = 0.2061Z_1 + 0.3187Z_2 + 0.4990Z_3 + 0.4819Z_4 + 0.6122Z_5 \tag{8}$$

$$F_3 = -0.0245Z_1 + 0.3812Z_2 + 0.8387Z_3 + 0.2243Z_4 + 0.3168Z_5 \tag{9}$$

The functions above are the principal component functions of environmental capacity ($F_1$), number of seeds ($F_2$), and root volume ($F_3$). For each of the Zi in the equations above, it

represents the standardized data of the subfactor $Z_i$, including capacity, number of seeds, distribution area, root volume, and score for the use value.

The specific equation to calculate the impact factor is as follow:

$$F = 0.6112F_1 + 0.2043F_2 + 0.0841F_3 \tag{10}$$

### 2.3. Analysis of Impact Factor

We select invasive species including Taraxacum officinale, Solidago canadensis in Wisconsin Madison, Pistia stratiotes in texas Austin, and other 23 distinct invasive species in some certain regions. Additionally, in our model, the higher the factor is, the more harmful the invasive species is.

For the dandelion (Taraxacum finale), the impact factor is 0.162, which is one of the lowest among all invasive species. The reason behind the result is that the spread of dandelions is quite limited–it could only spread in an open one-hectare of land. Moreover, the medical and edible values of this species are extremely high since it has a rich history of medicinal and culinary use.

**Table 1.** Impact factor

| Species | Impact Factor |
|---|---|
| Taraxacum finale | 0.162 |
| Pistia stratiotes | 2.493 |
| Eichhornia crassipes | 0.320 |

## 3. Prediction model of the spread of dandelions

### 3.1. Factors

Before evaluating the spread of the dandelion, we divide the influencing factors into two distinct categories: non-random factors and random factors.

### 3.1.1. Non-random Factors

For the random factors part, we mainly concentrate on the impacts of the wind speed and direction [4, 5]. For the wind speed, obviously, and speed is proportional to the spread distance. Furthermore, we have to consider not only the force inserted on the seeds and the air friction produced during the process, which means the net force should be as big as possible.

$$F_n = F_{wind} - F_d \tag{11}$$

$$F_d = 0.5\rho v^2 A C_d \tag{12}$$

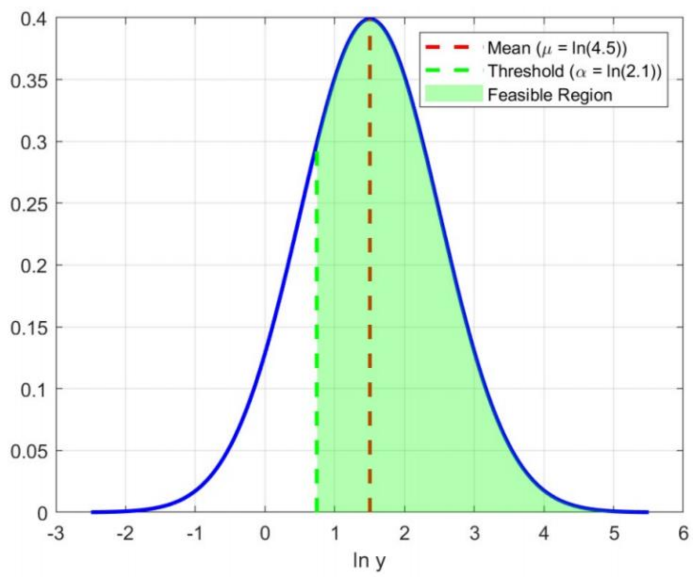


**Figure 1.** Normal Distribution of lnY

The dynamics and aerodynamics equations about the net force and the air drag, or air friction, mentioned above, we compute that equation and uncover that equation under the same circumstances, which means air density (ρ), the surface area of the seed (A), and drag coefficient ($C_d$) are the same, the more the speed of the wind is, the more the drag force is. As a result, a bigger speed can lead to less distance of the spread.

Then, we dig into the specific relationship. To illustrate, on a clear day, wind speeds can be regarded to have a lognormal distribution $\ln Y \sim N(\mu, \sigma^2)$. This statement indicates that the natural logarithm lnY is normally distributed with mean μ and variance $\sigma^2$ where $\sigma^2 = 1$. Figure 1 is the detailed normal distribution of lnY.

In order to interpret the data in a better way, we standardize the data, and the diagram above is the standardized normal distribution. Additionally, the area in green is the feasible region which means that the natural logarithm of the wind speed in that region can carry the seed to surrounding regions. So, we attain the mathematical formula of the normal distribution below.

$$lnY = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2\sigma^2}(lny-\mu)^2\right) & for\ y \geq 2.1 \\ 0 & for\ y < 2.1 \end{cases} \tag{13}$$

In the equation above, we define $\mu = \ln 4.5$ since$4.5\mathrm{m/s}$ is the best wind speed that could let the spread distance become the greatest distance. σ is the standard deviation of the wind speed.

### 3.1.2. Non-random Factors

According to the Euler's function, we represent the randomness of the spread angle as $e^{i\theta}$. In addition, θ in the formula represents the angle of the spread [6,7]. We determine to use a random number between 0.5 to 1 to calculate the random factors of the spread direction and distance. The below expression is the one we use to determine random factors.

$$e^{i\theta} rand[0.5,1] \tag{14}$$

After determining both the non-random factors and the random factors, we are able to summarize, or combine, the equation of the displacement of the seed.

$$x_{t+1} = x_t + e^{i\theta} rand[0.5,1] d_{maxs} \frac{\vec{y}}{\left|\vec{y}\right|} lnY \tag{15}$$

In equation above, xt represents the position of the seed at the time t. Moreover, $\frac{\vec{y}}{\left|\vec{y}\right|}$ is the unit vector of the wind, lnY is the normal distribution of the wind speed Y, and $d_{maxs}$ is the maximum spread distance in sunny days which is 16m. Thus, our equation of the spread is established.

## 3.2. The architecture of the multi-data prediction model

### 3.2.1. Survival Rate

After determining the optimized position of the spread of the seed above, we need to discover and consider the survival rate of the dandelion's seed since not all seeds will survive in each climate, and the rate change drastically under different scenario.

Like general plants, the growth rate, or the survival rate, of dandelion is mostly connected with temperature and humidity. Moreover, based on the expert's statement, the most suitable range of temperature is about 20°C to 25°C.The best range of humidity that dandelion has the greatest survival rate is 40% to 50%.

Then, we define an equation to compute the survival rate. Since, both the two extreme sides of temperature and humidity (extreme cold/hot/moist/drought) will result in lower rate of survival. The equation of the survival rate is as follow:

$$f(T,h) = rand[0.7,0.9] e^{-k_1(T-T_0)^2} e^{-k_2|h-h_0|} \tag{16}$$

In the equation above, $f(T,h)$ represents the survival rate of dandelions as a function of temperature and humidity. The term rand [0.7,0.9] denotes the germination rate of dandelion seeds, which is a random factor between 70% to 90% and is relatively unaffected by climate factors. The expression $e^{-k_1(T-T_0)}$ captures the exponential decline in survival rate as the temperature deviates from the optimal temperature, $T_0$ with $k_2$ being the temperature index that controls the sensitivity of survival rate to temperature changes. Similarly, $e^{-k_2|h-h_0|}$represents the impact of deviation from the optimal humidity $h_0$ on the survival rate, with $k_2$ as the humidity index. The values $T_0$=22.5°C and $h$0=45 are the best temperature and humidity for dandelion to thrive, and $k_2$=0.01 and $k_2$=1.29 are the determined indexes for temperature and humidity impact respectively.

To test whether the indexes ensure the principle that the change of survival rate in the rational and best range is small and maintained at a relatively higher level, we input several temperatures and humidity into our model, and the following chart is the survival rate in several different environmental conditions:

**Table 2.** Examples of survival rates in different environmental conditions

| Climate Type | Temperature (°C) | Humidity (%) | Survival Rate (%) |
|---|---|---|---|
| Mediterranean | 23.00 | 48.00 | 76.76 |
| Arid | 30.00 | 25.00 | 34.62 |
| Tropical | 30.00 | 60.00 | 36.93 |
| Temperate | 24.00 | 55.00 | 68.70 |

From the results and data above, we find that the indexes we set follows our requirements of the index. Therefore, the equation of indicating the survival rate in different condition is established.

### 3.2.2. Seed Density

Obviously, the capacity of the environment is not infinite, which means that even if some of the seeds are spread to other areas, they still do not have the chance to grow and spread their own seeds. Therefore, in order to make our predictions more rational, real, and logical, we define a maximum capacity to restrict the spread of seeds. Based on our observation, we uncover that the rational range of dandelions per square meter is from 40 to 70. Consequently, we randomly set the capacity from 40 to 70 dandelions per square meter. To be specific, if there were initially 30 dandelions, the random capacity of this square meter is 60, and 100 seeds were spread to this area, we think that the survival rate of the spread seed is about $\frac{60}{100+30}$. Therefore, we are able to define an equation of the seed density.

## 3.3. Results

### 3.3.1. Model Predictions

We collect the classical data of the climates we consider temperate, arid, tropical, and Mediterranean climates. After that, we compute the probability of a given temperature, humidity, and precipitation occurring on any given day of a given month. Then, we use it to simulate the weather conditions randomly according to the probability. The following diagrams and tables are the simulated results in distinct climates and months. The original position of the dandelion in the diagrams is (50, 50).

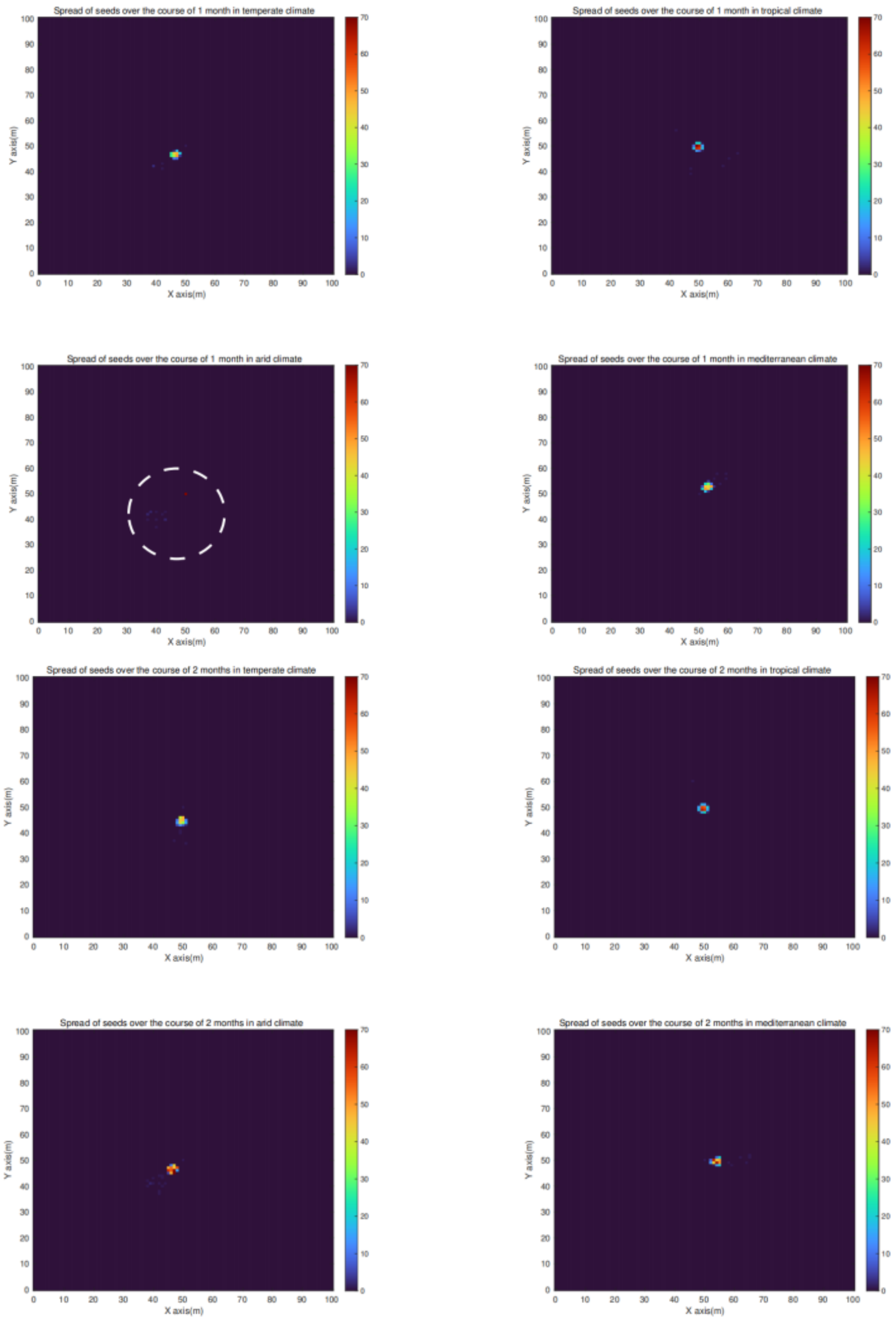


**Figure 2.** Spread of seeds over the course of 1 month and 2 month in different types of climate

**Table 3.** Average weather indexes of different climate in April.

| Climate | Temperature (°C) | Humidity (%) |
|---|---|---|
| Temperate | 19.25 | 45.45 |
| Tropical | 32.16 | 20.74 |
| Arid | 24.61 | 20.74 |
| Mediterranean | 15.18 | 59.30 |

**Table 4.** Average weather indexes of different climate in May.

| Climate | Temperature (°C) | Humidity (%) |
|---|---|---|
| Temperate | 23.97 | 49.44 |
| Tropical | 32.10 | 68.11 |
| Arid | 28.25 | 18.12 |
| Mediterranean | 20.44 | 54.98 |

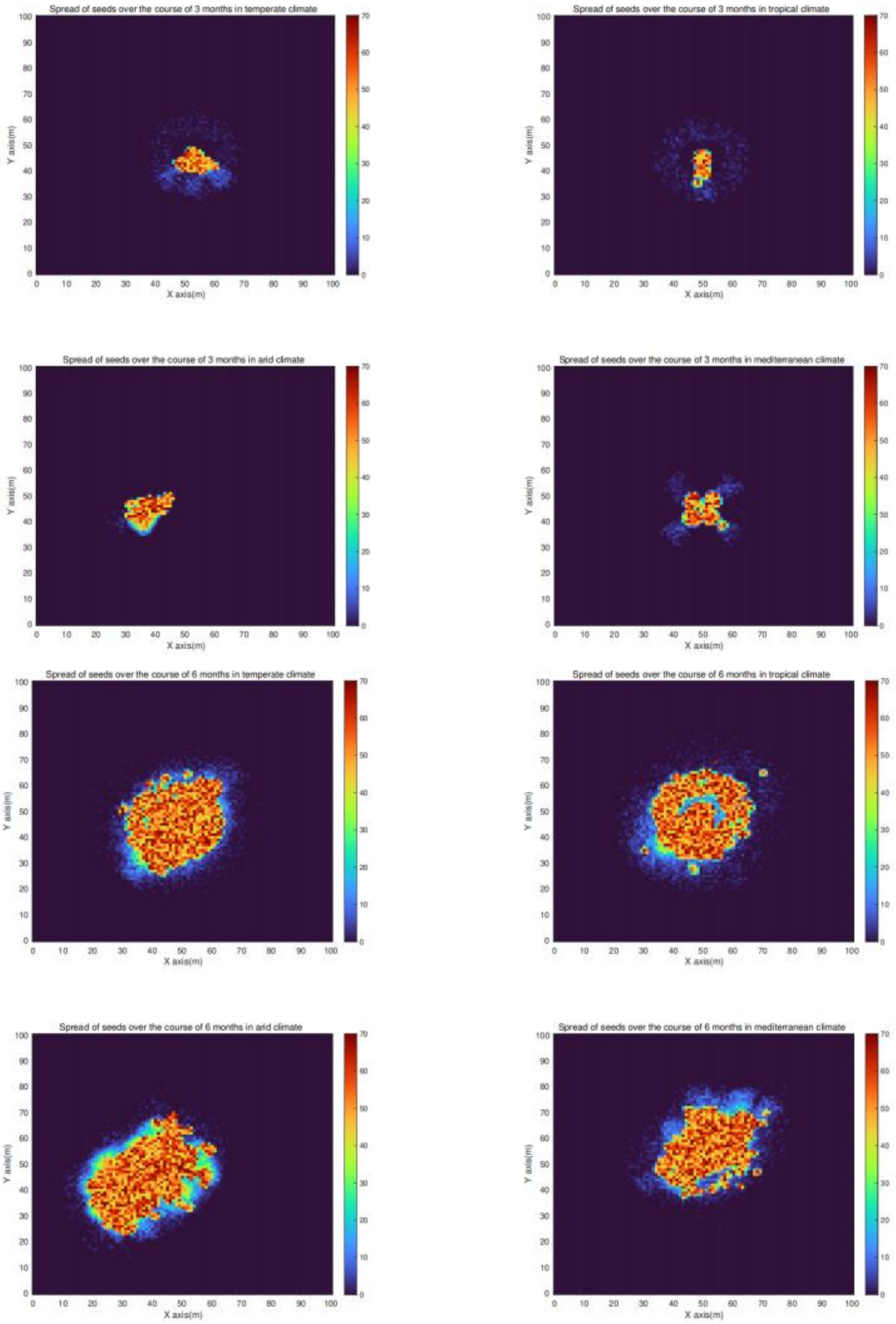


**Figure 3.** Spread of seeds over the course of 3 month and 6 month in different types of climate

**Table 5.** Average weather indexes of different climate in June.

| Climate | Temperature (°C) | Humidity (%) |
|---|---|---|
| Temperate | 29.06 | 54.50 |
| Tropical | 31.30 | 69.53 |
| Arid | 33.57 | 15.46 |
| Mediterranean | 24.66 | 52.62 |

**Table 6.** Average weather indexes of different climate in September.

| Climate | Temperature (°C) | Humidity (%) |
|---|---|---|
| Temperate | 16.52 | 60.91 |
| Tropical | 29.11 | 76.13 |
| Arid | 33.08 | 30.97 |
| Mediterranean | 23.81 | 51.19 |

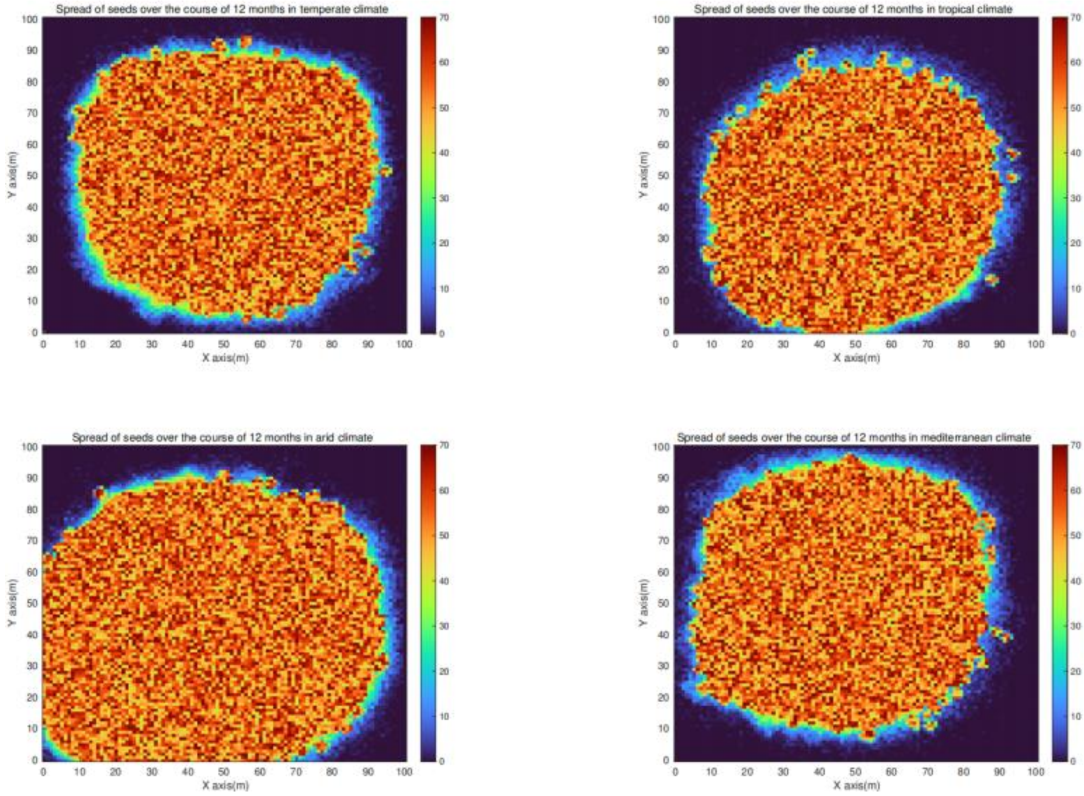


**Figure 4.** Spread of seeds over the course of 12 month in different types of climate

**Table 7.** Average weather indexes of different climate in March.

| Climate | Temperature (°C) | Humidity (%) |
|---|---|---|
| Temperate | 14.02 | 43.06 |
| Tropical | 31.45 | 68.75 |
| Arid | 18.52 | 37.03 |
| Mediterranean | 11.80 | 61.23 |

From the table and diagrams above, we can surprisingly be aware that the rough distributions of each type of climate are quite the same. We analyze the phenomenon and find the following conclusion: In a relatively long time, the differences in the seeds' distribution between climates will decrease. This is because even if some climates have the most suitable climate conditions enabling the spread to become optimized, they still have a large proportion of months that restrict the spread. For example, in temperate climates, spring and autumn are suitable for the spread; however, in the winter season, the cruel and cold conditions could decrease the survival rate and the spread distance. In a similar way, although all seasons of arid climate cannot let the spread become optimized, the condition can let the seeds spread at a relatively stable rate at any time of the year. Thus, the spread can consist all year long. All in all, after the trade-off between" optimizing" spread and" consisting" spread, the impacts of weather factors on the spread of dandelions will decrease over time.

### 3.3.2. Sensitivity Analysis

In order to examine how the setting of the maximum spread distance on sunny days and rainy days could influence the spread of dandelions, we changed the maximum spread distances $d_{maxs}$ and $d_{maxr}$ by $\pm 10\%$ and $\pm 20\%$ to analyze the sensitivity of the maximum spread distance in a temperate climate. We only consider the spread of dandelions over the course of 6 months and 12 months since the spread of dandelions at these times is larger and easy to observe differences.

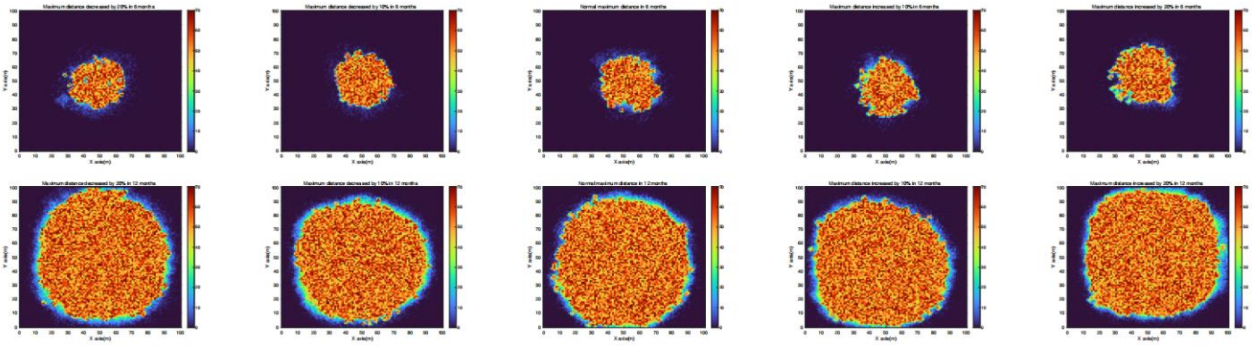

**Figure 5.** Sensitivity analysis of the maximum spread distance

The diagrams above show the changed maximum spread distance's impact on the spread of dandelions over the course of 6 months and 12 months. We can easily observe that despite the randomness of the wind direction, the spreads of dandelion in each maximum spread distance condition do not alter much. Moreover, because of the randomness of our model, in some cases, the range of spread of dandelions will be wider even if the maximum spread distance is decreased. Therefore, we believe that the maximum spread distance is not sensitive to the spread of dandelions.

## 4. Summary

In this paper, we utilize PCA to reduce dimensions by combining overlapping data, identifying the main components of impact factors to evaluate the influencing factors of dandelions, and compare them with other invasive species. In the dandelion spread prediction model established based on the DO algorithm, we take a comprehensive approach by not only considering the wind force normally but also factoring in the randomness of wind direction, the impacts of temperature and humidity on dispersion, aerodynamics, and the differences in spread on sunny versus rainy days. Through a comprehensive and integrated analysis, we achieve accurate predictions of the dandelion spread range.

However, in our prediction model, we have not deeply categorized climate types. In reality, there are several different subtypes within these general categories, each with its unique climate factors, which may lead to a lack of specificity in our model. Future research will further analyze the spread range predictions under more climate type conditions.

## References

[1] Seale M, Zhdanov O, Soons M B, et al. Environmental morphing enables informed dispersal of the dandelion diaspore [J]. Elife, 2022, 11: e81962.

[2] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, 374 (2065), 20150202.

[3] Zhao S, Zhang T, Ma S, et al. Dandelion Optimizer: A nature-inspired metaheuristic algorithm for engineering applications [J]. Engineering Applications of Artificial Intelligence, 2022, 114: 105075.

[4] Griffen, B. D., van den Akker, D., DiNuzzo, E. R., Anderson, L., & Vernier, A. (2021). Comparing methods for predicting the impacts of invasive species. Biological Invasions, 23, 491 - 505.

[5] Song, S., Shuang, S., Jiaxin, S., & Zhang, Y. (2016). Effects of different treatments on seed germination of dandelion. Ningxia Agriculture and Forestry Science and Technology, 57 (9), 5 - 6.

[6] Soons, M. B., & Bullock, J. M. (2008). Non-Random Seed Abscission Long-Distance Wind Dispersal and Plant Migration Rates. Journal of Ecology, 96 (4), 581 - 590.

[7] Tackenberg, O., Poschlod, P., & Kahmen, S. (2003). Dandelion Seed Dispersal: The Horizontal Wind Speed Does Not Matter for Long-Distance Dispersal – it is Updraft! Plant Biology, 5, 451 – 454.