2023

High School Mathematical Contest in Modeling (HiMCM)& Middle Mathematical Contest in Modeling (MidMCM)
Summary Sheet

Team Control Number: 14706
Problem Chosen: A

Dandelion diffusion distribution model and invasive species determination model

Dandelion is one of the most common plants in our lives. The common dandelion is known for its yellow flower heads that turn into round balls of many silver-tufted fruits that disperse in the wind.

In problem 1, our first thinking was to identify the influential factors for our model. We honed in on variables such as **humidity** and **temperature** as the most influential factors. We then delved into existing research to get more relevant data that would enable us to construct the **model of density curve** in the initial spread. To visualize the distribution of dandelions, we draw a random possibility curve of spread for the whole year. After arranging all the data, we can know that most of the seeds spread **2.2** meters away from the initial point. The distribution of seeds is **dense in the middle**, and **sparse around** them. While **0** in the 1st and 2nd months, the average density of seeds in the 3rd month is **15.9 seeds/m^2**, with the **0.15%** of coverage. **0.61%** of the area is covered by dandelions and the average density of seeds in the circle is **282.5 seeds/m$^{2.}$** In the 12th month, they are **4314577seeds/m^2** and **3.8%**.

In problem 2, we first use the conclusion of **cluster analysis** to construct the evaluation indicator system. Then we construct a **TOPSIS** model to determine the invasive degree of a random plant. We first took five sets of data of different plants. Then we input the data into a **TOPSIS** model, use the entropy method to finally got a formula. The larger the final score computed by the model, the lower the invasive level will the plant has. Since the final score of dandelion is higher than sunflower, a type of noninvasive species, we conclude that dandelion is **not invasive**. According to the conclusion, among the seven impact factors, the time to reproduce one generation has the greatest influence (**21.95%**), followed by maximum tolerance humidity (**16.23%**), minimum tolerance humidity (**15.89)** and other factors. The number of seeds per generation has the smallest influence at **6.6%**.

In general, we built models to predict dandelion wind spread pattern and determine whether it is an invasive plant. Through research and calculation, we take temperature and humidity as the main influencing factors and successfully draw the figures of wind-spread, summarize the spread pattern, and finally give specific results. When evaluating whether it is an invasive plant, we establish the TOPSIS model based on cluster analysis and compare it with other plants to draw a conclusion with the scores in the evaluation.

Key: spread pattern, humidity, temperature, model of density curve, model of invasive index prediction, TOPSIS, cluster analysis, entropy method.

Contents

1. Introduction

1.1 Background

Taraxacum officinale, the dandelion or common dandelion, is a perennial herbaceous flowering plant in the daisy family Asteraceae (syn. Compositae). The common dandelion is known for its yellow flower heads that turn into round balls of many silver-tufted fruits that disperse in the wind. Thus, dandelion is a common colonizer of disturbed habitats, both from wind-dispersed seeds and from seed bank germination. Seeds remain viable in the seed bank for many years, with one study showing germination after nine years. This species is a prolific seed producer, with 54 to 172 seeds produced per plant, and a single plant can produce more than 5,000 seeds per year. Besides, the plant has several culinary uses: the flowers are used to make dandelion wine, the greens are used in salads, the roots have been used to make a coffee substitute (when baked and ground into a powder), and the plant has been used by Native Americans as a food and medicine.

Nowadays, as a result of human economic activities and global interconnectedness, certain species have migrated from their natural habitats into foreign ecosystems, either through deliberate human action or by other means. In these new environments, they proliferate and form self-sustaining populations, and these organisms are categorized as invasive species. The intentional introduction of beneficial animal and plant species can enhance the biodiversity of the host country and provide various benefits. However, if introduced improperly or not managed properly, it can have significant negative consequences.

Figure 1. Dandelion

1.2 Problem Restatement

The key to this problem is to study the wind spread pattern of dandelion, and the problem can be divided into several small questions and steps:

Question 1： Establish a mathematical model to construct density curve and a random possibility curve with influencing factors of humidity and temperature. Calculate the average density and percent of coverage.

Question 2： Use the TOPSIS model based on cluster analysis to compare with other plants, determine the result and evaluate the influence factors by using indicator system established.

1.3 Our Work

In problem 1, first we focused on identifying influential factors for our model, such as humidity and temperature affecting dandelion distribution. These factors, being crucial, guided our efforts. Next, we recognized the importance of these factors in determining our model's outcome. This realization helped us pinpoint the key elements shaping our analysis. We then used existing research data to create a density curve in the initial spread, visually highlighting areas with the highest concentration of dandelions. Continuing our study, we aimed to understand the conditions promoting dandelion growth and spread, contributing to a comprehensive analysis. Finally, utilizing tools like spot plots, we shifted from observation to prediction, simulating the dandelion distribution over twelve months for a proactive approach to managing their spread.

In problem 2, we committed to find the model for the prediction of invasive plants. First, we found four plants' data, two invasive plants' and two non-invasive plants', we used the TOPSIS model to determine the differences between the data and then used the entropy method to give weights for the variables and finally got a formula. The larger the index that the formula processes, the lower the invasive level. Then, we applied the formula for the dandelion and got an index. If the index is larger than one of the non-invasive plants, this plant can be determined as a non-invasive plant. If the index is smaller than one of the invasive plants, this plant can be determined as an invasive plant.

2. Model Preparation

2.1 Assumptions

- *Assumption 1:* The number of seeds that dandelion can reproduce once is 400, but in the second question, to make it easier to calculate the comparison with the plant, this value is 200.

  Justification: Based on the reproductive ability of dandelion and various climatic conditions such as temperature, it is appropriate for us to take 400 as the result in problem 1. For the models we choose and steps we take in problem 2, 200 is a good choice.
- *Assumption 2:* The relationship between humility and sedimentation rate is 0.8.

  Justification: According to the research, there is a proportional relationship between humility and sedimentation rate, so we choose a proper ratio 0.8.

- *Assumption 3:* The relationship between temperature and number of surviving seeds is $Q = 400 - (T - 20)^2$

  Justification: When the temperature is near 20 degrees, the living condition of the seeds is the best, which causes the possible maximum output. Due to this reason, we assume the relationship to be this quadratic equation.
- *Assumption 4:* The low, medium, and high temperature range are 5-14 degrees, 15-24 degrees, and 25-35 degrees.

  Justification: According to the research, the temperature at which the seeds germinate is between 5 and 35 degrees, and we divide this into three temperature ranges.

2.1 Notations

The table 1 below defines all the variables we will use throughout this paper:

Table 1. Variable Chart

Symbol	Definition
μ	Wind speed
H	Seed release height
F	Seed settlement rate
ξ	Random variable of wind speed
ψ	Natural logarithm of ξ
$h(x)$	Probability density
$H(x)$	Distribution function
σ_x	Standard variance of random ψ
μ_x	Average value of random ψ
$RH(\%)$	Humidity
$E(y)$	Mathematical expectation
λ_m	Average distance of seeds by wind-dispersal
NS	Number of seeds per one generation
T	Time it takes for reproducing one generation
EI	Economic influence (in yuan per year)
LT	Lowest tolerance temperature (in celcius)
HT	Highest tolerance temperature (in celcius)
LH	Minimum tolerance humidity
HH	Maximum tolerance humidity
A	Spread area
r	Area of the circle during spread
n	Reproduction times
C	Percent of coverage
ρ	Density of seeds
N	Accumulative number of seeds
Q	Number of seeds produced in a single reproduction

3. Problem 1: Prediction of Dandelions Distribution

3.1 Analysis of problem

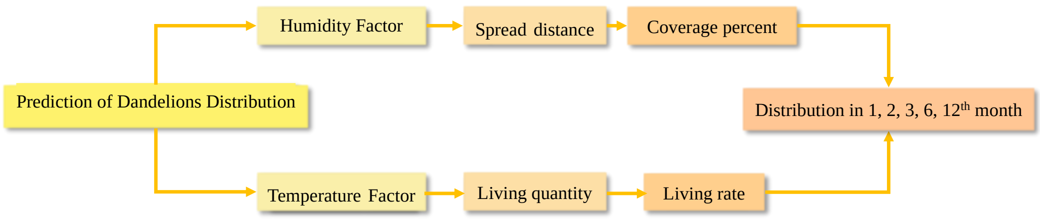

Figure 2. Flow chart for Problem 1

Dandelions are renowned as ubiquitous weeds, recognized for their vibrant yellow blossoms and unique, fluffy seed heads. Factors such as temperature, humidity, and wind velocity can further impact their prevalence and dispersal. Figure 2 is a flow chart for studying the distribution of dandelions.

3.2 Model of Dandelions Distribution

We found the information and formulas from the research [1].

We first consider the influence of wind. Consulting the research, the relationship between seed wind spread distance x and wind speed μ is

$$x = \frac{\mu H}{F}. \tag{1}$$

H is the seed release height, and F is the seed settlement rate.

Let ψ be the natural logarithm of the random variable ξ of wind speed, that is, $\psi = In\xi,$ whose probability density is $h(x)$ and distribution function is $H(x)$. According to the definition and properties of the probability distribution:

$$F(y) = P(\xi \le y) = P\left(e^{\Psi} \le y\right) = P(\psi \le \ln y) = \int_0^{\ln y} h(x)dx. \tag{2}$$

Let the probability distribution of wind speed be lognormal. The probability density function of ψ:

$$h(x) = \frac{1}{\sigma_x \cdot \sqrt{2\pi}} \cdot \exp\left(-\left[\frac{x-\mu_x}{\sqrt{2}\sigma_x}\right]^2\right). \tag{3}$$

σ_x is the standard variance of random variable ψ, and μ_x is the average value of random variable ψ.

According to the characteristics of the random variable function, the wind speed probability density function is

$$f(y)=\frac{1}{y\sigma_{\ln y}\sqrt{2\pi}}\cdot\exp\left(-\left[\frac{\ln y-\mu_{\ln y}}{\sqrt{2}\sigma_{\ln y}}\right]^2\right). \tag{4}$$

and

$$\mu_{\ln y}=\ln\left(\prod_{i=1}^{n}y_i^{\frac{1}{n}}\right). \tag{5}$$

From (1.4), according to the solution of maximum value of the function, we have:

$$y=\frac{\exp\left(\mu_{\ln y}\right)}{\exp\left(\sigma_{\ln y}^2\right)}. \tag{6}$$

The mathematical expectation of the random variable of wind speed in (1.4) is

$$E(y)=\exp\left(\mu_{\ln y}+0.5\sigma_{\ln y}^2\right). \tag{7}$$

Plug (7) into (1), we get that the average distance of seed by wind-spread is

$$\lambda_m=\frac{E(y)H}{F}. \tag{8}$$

We assume that the number of seeds propagated in each direction is roughly the same, and analyze the profile of seed propagation. Using the data from the research [1] we assume $RH = 0.64$, $Q = 200$, $F = 0.8 \cdot RH$, $H = 0.400$ and predict the following density curve:

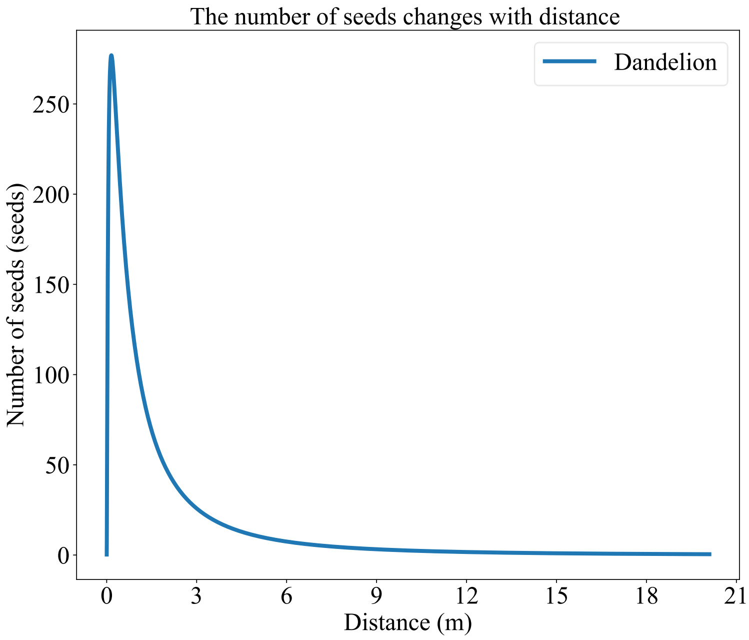

Figure 3. The number of seeds changes with distance

From figure 3, density is determined by the number of seeds so that we can find the location with the most dandelions is approximately X meters away from the initial dandelion. Between the densest point and the starting point, the concentration of dandelions increases as the distance from the origin increases. Beyond the densest circle, the dandelion density decreases with increasing distance, approaching zero.

3.3 Influences by Factors

4.3.1 Humidity Factor

Dandelion is an herbaceous plant that can spread with the wind. It often changes the distance between the seed landing site and the mother plant with changes in wind speed, humidity, height, and other factors. Due to various factors, according to the research, we obtained an average distance. Based on this average distance, we used a mathematical model to simulate the area of dispersal of dandelion.

If the number of the seeds each time one dandelion can produce is 200 to see the seeds more clearly. We analyze the profile of seed propagation. Using the data from the research[1] as well, we assume $RH = 0.64$, $Q = 200$, $F = 0.8 \cdot RH$, $H = 0.400$and predict the following predicted graph, simulated by the model we created:

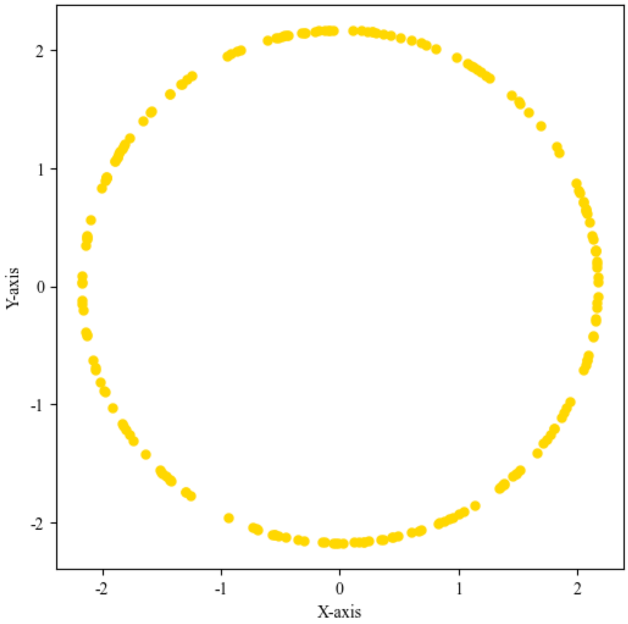

Figure 4. Predicted graph

The graph shows the first production of the dandelion. Dandelion is distributed from the center outward on a land of 2.2m * 2.2m averagely.

To further simulate the distribution of the dandelions for a longer time, we made the second reproduction process's graph and the third one. Following graphs shows the distribution in the second and third period, if the number of seeds produced is 1000, in order to make the graph more pleasing to the eye.

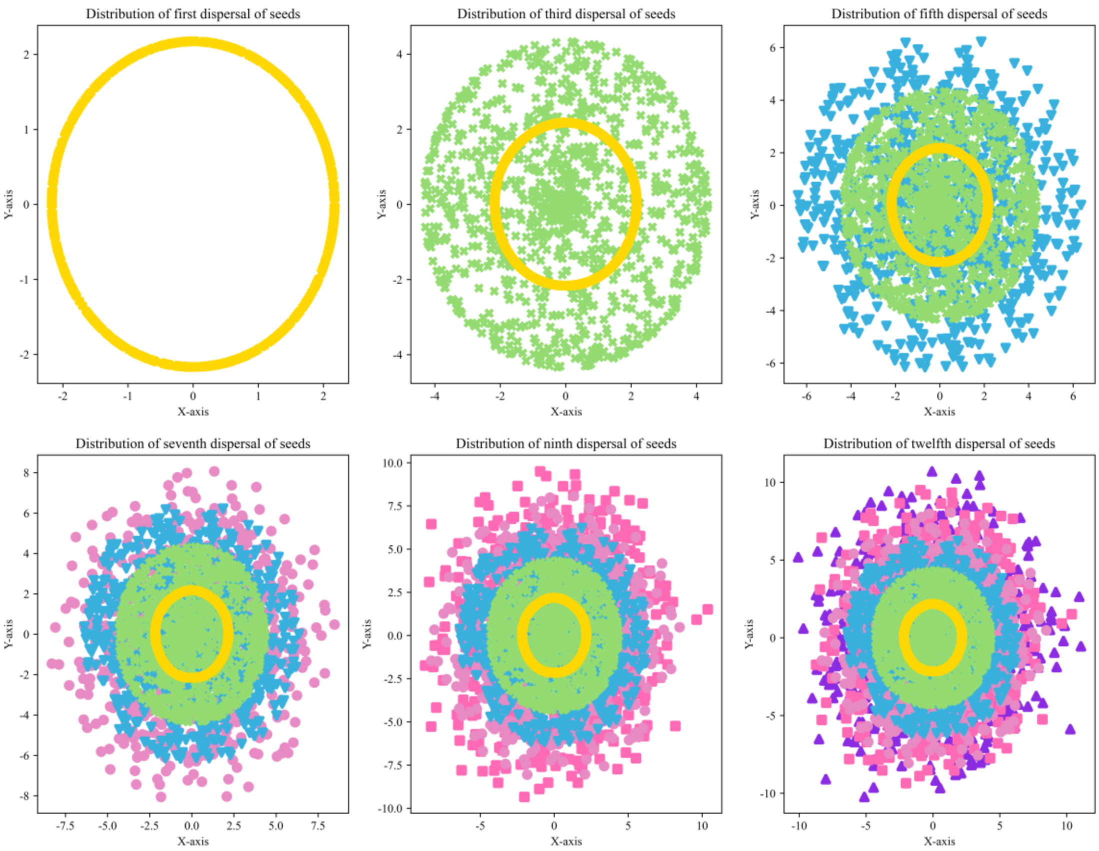

Figure 5. Distribution of seed dispersal in different month

In the figure 7, each different month corresponds to a different color. The first month corresponds to the center of the image. Dandelion breeds every two months, displaying February, April, June, August, October, and December, while the other months are the same as the previous reproduction image. The first month is a dandelion plant, and after two months, the first production is a golden circle with the most seeds at a time. The fourth month corresponds to the purple part in the picture, which is still rich and dense, but not as dense as February. From then on, the density and quantity decreased month by month until December and spread outward.

Estimating the graph, the average distance each spread is 2.2, so after times of production and spread, we can know that from January to December, the radius of the spread circle is 13.2 meters long. Then, we can use the formula (9) for calculating the circle's area.

$$A = \pi r^2 \tag{9}$$

r is the radius of the spread circle. A is the area of the circle. A represents the spread area, while r represents radius. In the situation, the radius is:

$$r = 2.2^n \tag{10}$$

n is the times that dandelions reproduce seeds.

Finally, considering that the total area is one hectare, 100m*100m, we can calculate the percent of coverage, using following formula:

$$C = \frac{A}{100 \times 100} \tag{11}$$

Let C represents the percent of coverage in 1 hectare.

Next, to better understand dandelion growth, we need to calculate the average density of the area per month, using the following formula:

$$\rho = \frac{N}{A} \tag{12}$$

ρ is the density of seeds in the present area, while N represents the accumulative number of seeds. To find the number of accumulated seeds, we use the exponential function:

$$N = Q^n \tag{13}$$

We will take the humidity factor into account that has influenced the distribution of dandelions in seed dispersion. Referring to the research[2], parachute opening is modulated by the level of humidity in the atmosphere: higher humidity triggered swelling in the actuator and mechanical movement of hairs upwards, which closes the parachutes and increases the sedimentation rate. As a result, there is a proportional relationship between humidity and sedimentation rate.

Therefore, we made our assumption. We assumed that the relationship between humidity and sedimentation rate is:

$$F = 0.8RH\,. \tag{14}$$

where RH refers to the humidity, and F denotes seed settlement rate.

In order to study the influence by humidity, we use three sets of data to draw the density curve, related to the distance from the initial point.

Table 2. Data set 1 of humidity

Parameter	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
RH (%)	0.730	0.604	0.541	0.596	0.608	0.614
H (m)	0.40	0.20	0.30	0.25	0.15	0.25

We will take the temperature factor into account that has influenced the distribution of dandelions in seed dispersion. Referring to the research [3], seeds only germinate when the temperature is between 5-30 degrees Celsius, but the germination rate varies. Between 5-14 degrees Celsius, the germination rate is the lowest. The highest germination rate occurs between 15-19 degrees Celsius. However, above 20 degrees Celsius, the seed germination rate decreases. The number of living seeds is modulated by the temperature: As a result, there is a proportional relationship between humidity and sedimentation rate.

$$Q = 400 - (T - 20)^2 \tag{15}$$

Q is the number of living seeds spread; T is temperature. Following table is the data we get from the research (5):

Table 3. Data set 2 of temperature

Parameter	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
T (℃)	8	15	25	28	30	35
H (m)	0.4	0.4	0.4	0.4	0.4	0.4

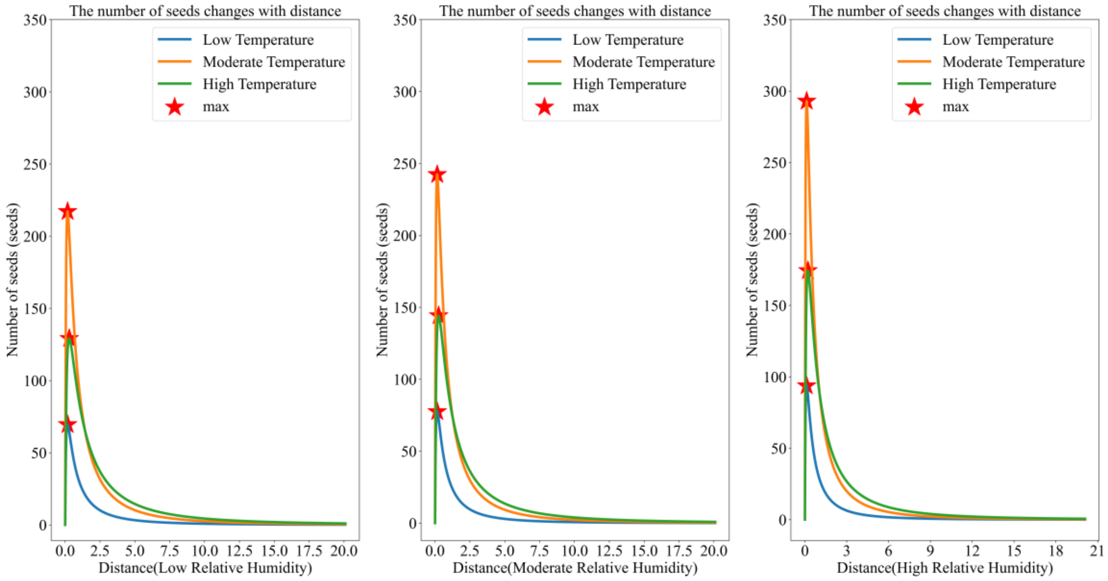

Figure 4. The number of seeds changes with distance in different temperature

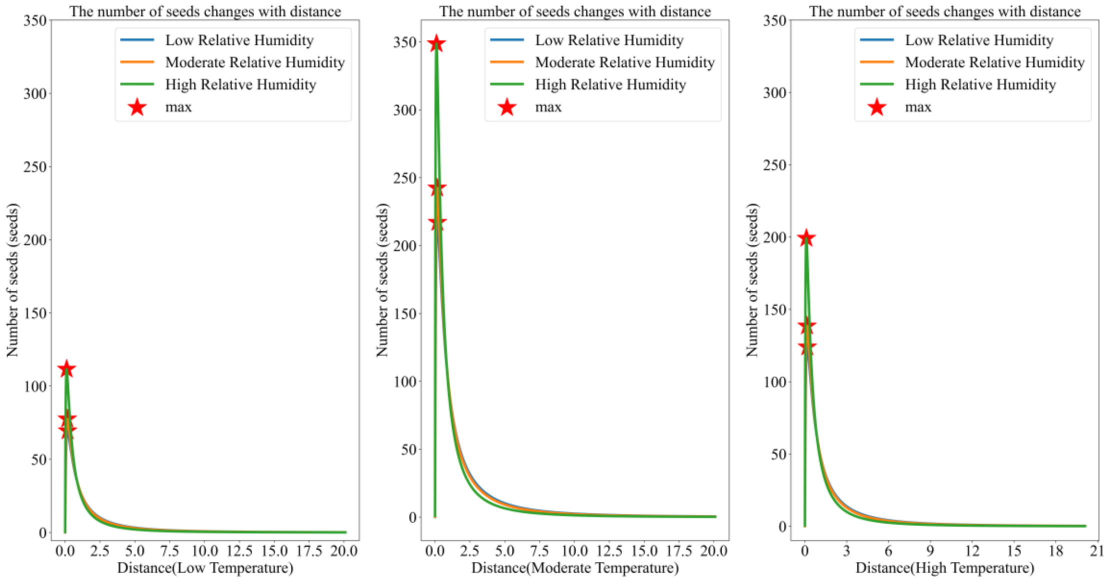

Figure 6. The number of seeds changes with distance in different humidity

From figure 4 and 5, we can clearly see that humidity has an impact on the quantity of seeds in the densest areas: the higher the relative humidity, the more dandelions grow in relatively dense areas. Additionally, humidity also affects the distance of seed dispersal: the higher the relative humidity, the fewer dandelions grow in relatively remote areas.

Table 4. Maximum amount of seeds in different humidity and temperature

Max(seeds) T(℃) RH(%)	8	15	25
0.730	292.9	292.9	174.3
0.604	144.2	242.3	77.5
0.541	129.2	217.1	69.5

Through the curve, we can determine that under moderate temperature conditions, the highest number of seeds survive, followed by high temperatures, and the lowest survival rate occurs under low temperatures; we can determine that under moderate temperature conditions, the highest number of seeds survive, followed by high temperatures, and the lowest survival rate occurs under low temperatures.

3.4 Model Results and Analysis

The more moderate the temperature, the more seeds survive, the higher the humidity, and the smaller the area of propagation. As can be seen from the density curve, in a dry and moderate temperature environment, the densest part of the seeds is far from the origin and the number of seeds in this area is greater than in other situations. However, high, or low temperatures can cause a decrease in the number of seeds, resulting in a decrease in the number of seeds in the densest part of propagation. In humid conditions, the dense areas of seed propagation are close together.

After that, the percent of coverage and the density of seeds in the present area and shown in the table:

Table 5. Percent of coverage and the density of seeds

Humidity (%) and Temperature (℃) Density & Percent of Coverage (%) Time(month)	RH=0.730 T=8		RH=0.730 T=15		RH=0.730 T=25		RH=0.604 T=8	
1	N/A	0.00E+00	N/A	0.00E+00	N/A	0.00E+00	N/A	0.00E+00
2	N/A	0.00E+00	N/A	0.00E+00	N/A	0.00E+00	N/A	0.00E+00
3	1.15E+01	1.50E-01	1.93E+01	1.50E-01	1.15E+01	1.50E-01	9.50E+00	1.50E-01
6	4.13E+02	6.10E-01	1.17E+03	6.10E-01	4.22E+02	6.10E-01	2.83E+02	6.10E-01
12	1.93E+07	3.80E+00	2.58E+08	3.80E+00	1.93E+10	3.80E+00	4.31E+06	3.80E+00

RH=0.604 T=15		RH=0.604 T=25		RH=0.541 T=8		RH=0.541 T=15		RH=0.541 T=25	
N/A	0.00E+00	N/A	0.00E+00	N/A	0.00E+00	N/A	0.00E+00	N/A	0.00E+00
N/A	0.00E+00	N/A	0.00E+00	N/A	0.00E+00	N/A	0.00E+00	N/A	0.00E+00
1.59E+01	1.50E-01	5.10E+00	1.50E-01	8.50E+00	1.50E-01	2.38E+00	1.92E+00	1.45E+00	9.84E-01
7.98E+02	6.10E-01	8.16E+01	6.10E-01	2.27E+02	6.10E-01	6.40E+02	6.10E-01	6.56E+01	6.10E-01
1.94E+05	3.80E+00	3.35E+05	3.80E+00	4.31E+06	3.80E+00	5.78E+07	3.80E+00	1.94E+05	3.80E+00

From the table, we can see that in the 1st month, only one seed is in the field, so there is no area covered and the same for the 2nd month. In the 3rd month, the dandelion begins to spread. From then on, in various conditions, the percent of coverage and average density differentiate from each other. Take RH=0.604, T=15 as an example to explain. In the 3rd month, the coverage of dandelions is 0.15% of the total area. Moreover, the average density of seeds in the spread circle is 15.9seeds/m^2. Then, in the 6th month, the month is not a mature month for the dandelions to spread. In that case, the average density of dandelions is 282.5seeds/m^2 and the coverage is 0.61%. Finally, in the 12th month, like 6th month, it is not a mature month, so the distribution and spread are like in the 11th month. The percent of coverage is 3.8%, with the average density of 4314577seeds/m^2.

4. Problem 2: Model for determining invasive plants and their "impact factor"

4.1 Selection of Evaluation Indicators

4.1.1 Using cluster analysis to determine the influence factors

We use cluster analysis (k-means) to help us choose the variables that determine the invasive index of the plants.

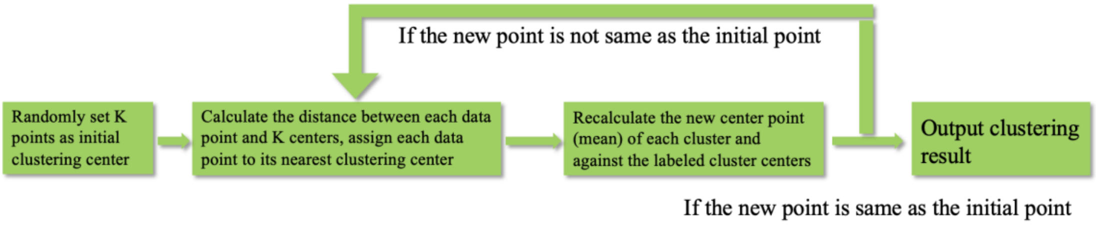
If the new point is not same as the initial point
Randomly set K points as initial clustering center
Calculate the distance between each data point and K centers, assign each data point to its nearest clustering center
Recalculate the new center point (mean) of each cluster and against the labeled cluster centers
Output clustering result
If the new point is same as the initial point

Figure 7. Flow chart for cluster analysis

- Step 1: Randomly set K points in the feature space as the initial clustering center.

- Step 2: The distance between each data point and K centers is computed, and each unclassified data point is assigned to its nearest cluster center point. One cluster represents one category. The formula for calculating distance is as follows:

$$d=\sqrt{\left(x_n-x_1\right)^2+\left(y_n-y_1\right)^2}\,. \tag{16}$$

In cluster analysis, we considered different factors of each plant species, including: edibility, average growth rate, reproduction speed, toxicity, ability to compete for nutrients with other plants, amount of oxygen produced, medicinal properties, ornamental value, seed production speed, dispersal ability, extensive root system and environmental tolerance of each plant species.

Table 6 shows the data we used in cluster analysis. Due to space limitation, only part of the data is shown:

Table 6. Data set 1 for cluster analysis

Species	**Height Average Minimum (cm)**	**Height Average Maximum (cm)**	**Category**	**Edibility**	**Average Growth Rate**	**Reproduction Speed**
Bishop's weed	30	100	L	Y	F	F
Cheatgrass	30	90	W	N	F	F
Asiatic sand sedge	20	40	L	N	M	M

In order to facilitate the subsequent calculation, we convert all qualitative indicators into numerical values. Table 7 shows the meaning and value of each notation:

Table 7. Meaning and value of each notation

Notation	**Meaning**	**Values**
L	Species that are present but localized	1
W	Species that are wide-spreading	0
Y	Yes	1
N	No	0
F	Fast	2
M	Medium	1
S	Slow	0
H	High	1
L	Low	0

Then we combine the contents of Table 6 and Table 7 to get Table 8：

Table 8. Data set 2 for cluster analysis

Species	Height Average Minimum (cm)	Height Average Maximum(cm)	Category	Edibility	Average Growth Rate	Reproduction Speed
Bishop's weed	30	100	1	1	2	2
Cheatgrass	30	90	0	0	2	2
Asiatic sand sedge	20	40	1	0	1	1

- Step 3: The new center point (mean) of each cluster is then recalculated against the labeled cluster centers. Formulas are as follows:

$$P_y = \sum_{i=1}^{n} p_{iy} \Big/ n. \tag{17}$$

$$P_x = \sum_{i=1}^{n} p_{ix} \Big/ n. \tag{18}$$

- Step 4: If the calculated new center point is the same as the original center point (the center of data no longer moves), then end the process, otherwise repeat step two.
- Step 5: When the result of each iteration is unchanged, the algorithm is considered to converge and the clustering is complete.
- Step 6: model analysis with elbow method. The graph is used to select the best number of clusters. The abscess is the number of clusters and the ordinate is K-means. The loss function of a cluster is the sum of the squares of the distances of all samples to the center of the class, which is the sum of the squares of the errors (the smaller the value, the better the clustering effect).

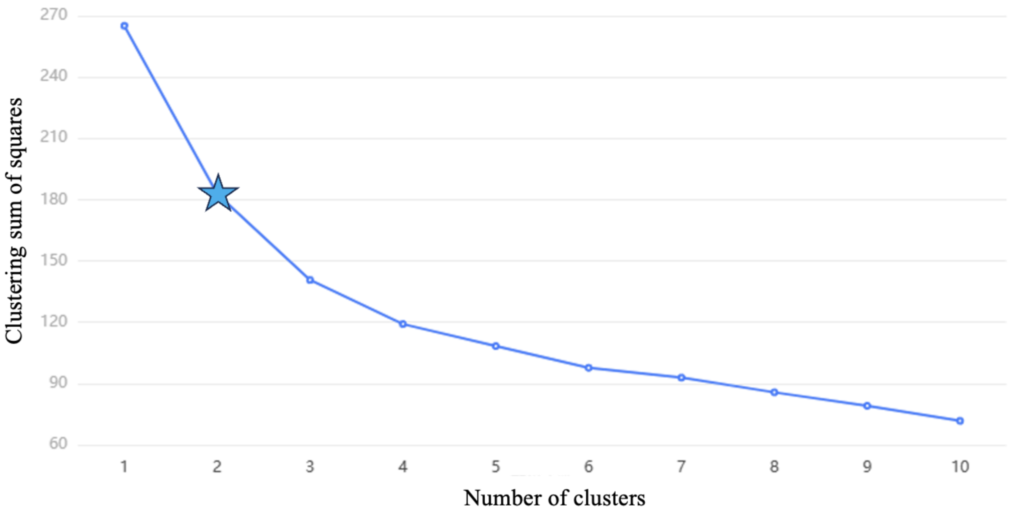

Figure 8. Elbow graph

In the elbow graph (figure 9), we found that the gradient difference is biggest when x=2 which represent the best number for clusters is 2 which represent invasive plants group and non-invasive plants group. It also proves our variables' validity.

Table 9. Result of cluster analysis

	Clustering category（average ± standard deviation）		F	P
	category1(n=43)	category2(n=33)		
Edible	1.279±0.504	1.121±0.331	2.43	0.123
Average Growth Rate	0.0±0.0	0.818±0.392	188.408	0.000***
reproduction speed	0.163±0.374	1.0±0.25	123.203	0.000***
Is it toxic	0.07±0.258	0.061±0.242	0.025	0.875
Robbing other plants of nutrients	0.907±0.294	0.818±0.392	1.276	0.262
Amount of Oxygen produced	1.116±0.981	1.455±0.617	3.006	0.087*
Medically useful	0.326±0.474	0.273±0.452	0.241	0.625
Is it ornamental	0.698±0.465	0.758±0.435	0.328	0.569
seed production speed	0.186±0.394	0.939±0.348	75.467	0.000***
dispersal ability	0.279±0.454	1.061±0.242	80.144	0.000***
Massive root	0.767±0.427	0.121±0.331	51.566	0.000***
Environmental tolerance	0.047±0.305	1.182±0.846	66.41	0.000***

Note: ***, ** and * represent significance levels of 1%, 5% and 10% respectively.

In the table, there are p values for each variable which represent the significance of the variable. (The larger the significance, the smaller the value of using this variable to determine the invasive index) When $p < 0.05$, the variable is significant and rejects null hypothesis, otherwise, the variable is accepted. Through the analysis, we found there are 7 variables that are accepted which are **Average Growth Rate, reproduction speed, Amount of Oxygen produced, seed production speed, dispersal ability, massive root and Environmental tolerance.**

4.1.2 Construction of indicator system based on the conclusion of cluster analysis

Based on the conclusion draw by cluster analysis, we conclude that the capacity of reproduction, economic impact the species can bring, and adaptability of the species to the environment are the three key factors in determining the invasive species.

Based on this definition, our essay selects the seeds per generation of the investigated species, the time required to produce a generation, the economic impact, and the range of environmental extremes to which the species is adapted, including the maximum and minimum temperatures and humidity, as the comprehensive evaluation indexes in order to evaluate the degree of invasion of a species.

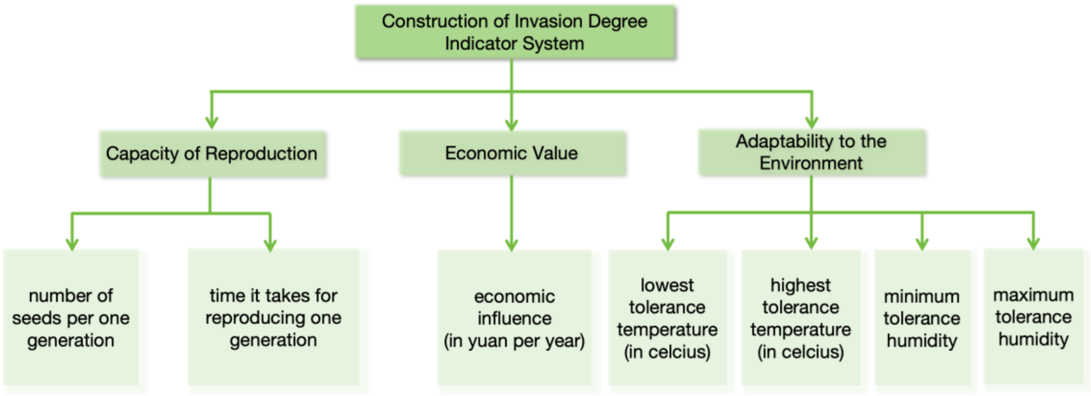
Construction of Invasion Degree Indicator System
Capacity of Reproduction
Economic Value
Adaptability to the Environment
number of seeds per one generation
time it takes for reproducing one generation
economic influence (in yuan per year)
lowest tolerance temperature (in celcius)
highest tolerance temperature (in celcius)
minimum tolerance humidity
maximum tolerance humidity

Figure 9. Construction of invasion degree indicator system

The following table shows the data we used in Topsis. Due to space limitation, only part of the data is shown:

Table 10. Data of each species

Species	**NS**	**T**	**EI** (in yuan per year)	**LT**	**HT**	**LH**	**HH**
Tumbleweed	230,000	2.5	-2.2E+10	-2.2	43.3	0.10	0.30
Dandelion	2,000	3	2.0E+09	5	30	0.50	0.80
Canada Goldenrod	22,500	7	-2.4E+10	-5	40	0.10	0.70
Sunflower	2,000	2.5	3.6E+10	4	44	0.50	0.70
Osmanthus Fragrans	30	5.5	2.1E+09	-13	33	0.75	0.85

4.2 Introduction of Entropy-weighted TOPSIS

A comprehensive evaluation model based on entropy-weighted TOPSIS method is established. The TOPSIS method is based on normalizing the original data matrix, forming a space between the best and worst solutions in the limited scheme, and then considering the object to be evaluated as a point in space. The distance between the object and the best and worst solutions is measured, and then the distance between different evaluation objects and the best solution is compared to obtain the order of superiority and inferiority. Objective determination of the weight of each index and the degree of invasion of the chosen species could be obtained based on the entropy method.

4.3 The Establishment and Solution of TOPSIS Model

- Step 1: Positivization of Evaluation Indicators

The "Benefit type" indicator is the indicator that its value is the bigger the better, the "Cost type" indicator is the indicator that its value is the smaller the better. Among the seven evaluation indicators selected above, time needed for reproducing one generation, economic impact, and minimum tolerance temperature and humidity are benefit type indicators; seeds per generation and maximum adapted temperature and humidity are cost type indicators. In order to eliminate the influence of different properties between types of indicators and facilitate the calculation of subsequent research, we positivize the indicators. Since the selected indicators only include benefit type and cost type indicators, we just need to convert the latter one into benefit type indicator. The equation for positivization is shown in Equation 1.

Cost type indicator to benefit type：

$$\max - x \ . \tag{19}$$

- Step 2: Standardization of Indicators

The invasion degree evaluation matrix X_ij(i=1, 2,……n; j =1, 2,……m; n=5, m=7) is constructed after positivization. This matrix is standardized to remove the influence of different units and the standardized matrix is noted as Z.

Standardization formula：

$$Z_{ij} = \frac{X_{ij}}{\sqrt{\sum_{i=1}^{n} X_{ij}^{2}}} \ . \tag{20}$$

From this formula, we get standardized matrix Z：

$$\begin{bmatrix} 0.0000 & 0.2491 & -0.4564 & -0.1421 & 0.0383 & 0.0961 & 0.9297 \\ 0.5099 & 0.2989 & 0.0415 & 0.3229 & 0.7666 & 0.4806 & 0.0845 \\ 0.4641 & 0.6974 & -0.4927 & -0.3229 & 0.2190 & 0.0961 & 0.2535 \\ 0.5099 & 0.2491 & 0.7385 & 0.2583 & 0.0000 & 0.4806 & 0.2535 \\ 0.5143 & 0.5479 & 0.0436 & -0.8394 & 0.6024 & 0.7209 & 0.0000 \end{bmatrix}$$

Since there are negative numbers in the Z matrix obtained from the previous standardization, X needs to be re-standardized, and the specific formula is as follows：

$$Z_{ij} = \frac{X_{ij} - \min\left\{x_{1j}, \dots, x_{nj}\right\}}{\max\left\{x_{1j}, \dots, x_{nj}\right\} - \min\left\{x_{1j}, \dots, x_{nj}\right\}} \ . \tag{21}$$

The re-standardized matrix Z obtained by the formula above is:

$$\begin{bmatrix} 0.0000 & 0.0000 & 0.0295 & 0.6000 & 0.0500 & 0.0000 & 1.0000 \\ 0.9914 & 0.1111 & 0.4339 & 1.0000 & 1.0000 & 0.6154 & 0.0909 \\ 0.9023 & 1.0000 & 0.0000 & 0.4444 & 0.2857 & 0.0000 & 0.2727 \\ 0.9914 & 0.0000 & 1.0000 & 0.9444 & 0.0000 & 0.6154 & 0.2727 \\ 1.0000 & 0.6667 & 0.4356 & 0.0000 & 0.7857 & 1.0000 & 0.0000 \end{bmatrix}$$

- Step 3: Calculation of Weight Based on Entropy Method

We calculate the probability matrix P based on the re-standardized non-negative matrix Z. Each element in P can be obtained by formula (2.7). Next, we calculate the entropy value of each indicator according to formula (2.8) and the information utility value using formula (2.9). The entropy weight of each indicator is obtained after normalization according to equation (2.10).

The formulas are as follows：

$$p_{ij} = \frac{z_{ij}}{\sum_{i=1}^{n} z_{ij}}. \tag{22}$$

$$e_j = -\frac{1}{\ln(n)} \sum_{i=1}^{n} \left[p_{ij} \cdot \ln(p_{ij}) \right], j = \{1, 2, 3, \ldots m\}. \tag{23}$$

$$d_j = 1 - e_j \tag{24}$$

$$W_j = \frac{d_j}{\sum_{j=1}^{m} d_j}, j = \{1, 2, 3, \ldots, m\}. \tag{25}$$

The following table shows the weight calculation results of each indicator：

Table 11. Weight calculation results by entropy method

Indicators	e_j	d_j	W_j（%）
NS	0.8608	0.1392	6.60%
T	0.5373	0.4627	21.95%
EI	0.6694	0.3306	15.68%
LT	0.8302	0.1698	8.06%
HT	0.6715	0.3285	15.58%
LH	0.665	0.335	15.89%
HH	0.6579	0.3421	16.23%

The entropy value reflects the degree of dispersion of the indicator, the larger the entropy value, the smaller the degree of dispersion of the indicator, the smaller the importance of the indicator for decision-making. Conversely, the smaller the entropy value, the greater the degree of dispersion of the indicator, the more information it provides, the more significant its role in the comprehensive evaluation, hence the greater its weight. The above and following charts show that the order of weights from the biggest to the smallest one is 1. time it takes for reproducing one generation, 2. maximum tolerance humidity, 3. minimum tolerance humidity, 4. economic influence, 5. highest tolerance temperature, 6. lowest tolerance temperature, 7. number of seeds per one generation. From this sequence we can conclude that the time it takes for creating one generation has the greatest influence on determining the degree of invasion, and the number of seeds per one generation has the smallest influence.

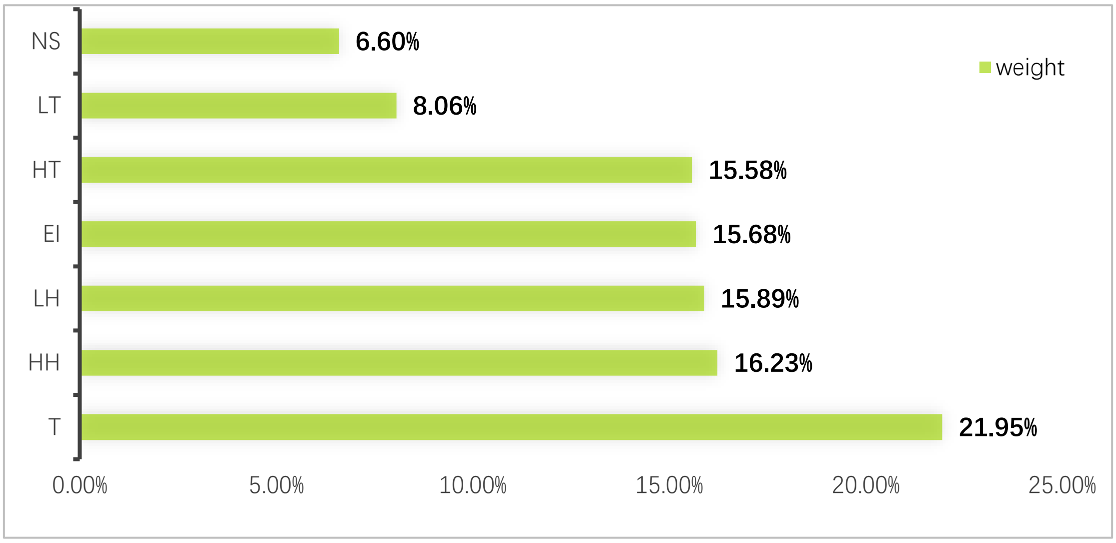

Figure 10. Weight calculation results by entropy method

- Step 4: Calculation of the final score

We take the maximum value of each indicator as the best value and the minimum value as the worst value, as shown in formulas (2.11) and (2.12). Then according to formula (2.13) and (2.14), the distance between the i th (i=1, 2, ... n) evaluation object and the maximum and minimum values is calculated respectively. Calculate the score of each evaluation object based on formula (2.15) and then normalize those scores according to formula (2.16).

$$Z^+ = (Z_1^+, Z_2^+, ..., Z_m^+) = \{\max z_{i1}, \max z_{i2}, ..., \max z_{im}\}. \tag{26}$$

$$Z^- = (Z_1^-, Z_2^-, ..., Z_m^-) = \{\min z_{i1}, \min z_{i2}, ..., \min z_{im}\}. \tag{27}$$

$$D_i^+ = \sqrt{\sum_{j=1}^{m} W_j \left(Z_i^+ - Z_{ij}\right)^2}. \tag{28}$$

$$D_i^- = \sqrt{\sum_{j=1}^{m} W_j \left(Z_i^- - Z_{ij}\right)^2}. \tag{29}$$

$$S_i = \frac{D_i^-}{D_i^+ + D_i^-}. \tag{30}$$

$$\tilde{S}_i = \frac{S_i}{\sum_{i=1}^{n} S_i}. \tag{31}$$

The larger the D^- value is， the greater the distance between the indicator and the worst solution, the better the indicator is. The larger the $\widetilde{S_i}$ value is, the better the evaluation object is, the less invasive a species is.

The calculation and sortied results are as follows：

Table 12. Final scores and ranking results of each species

Species	D_i^-	D_i^+	S_i	$\widetilde{S_i}$	Ranking results
Tumbleweed	0.438	0.8635	0.3365	0.1472	5
Dandelion	0.6285	0.6175	0.5044	0.2206	2
Canada Goldenrod	0.5603	0.7118	0.4405	0.1926	4
Sunflower	0.6049	0.6964	0.4648	0.2033	3
Osmanthus Fragrans	0.6697	0.5695	0.5404	0.2363	1

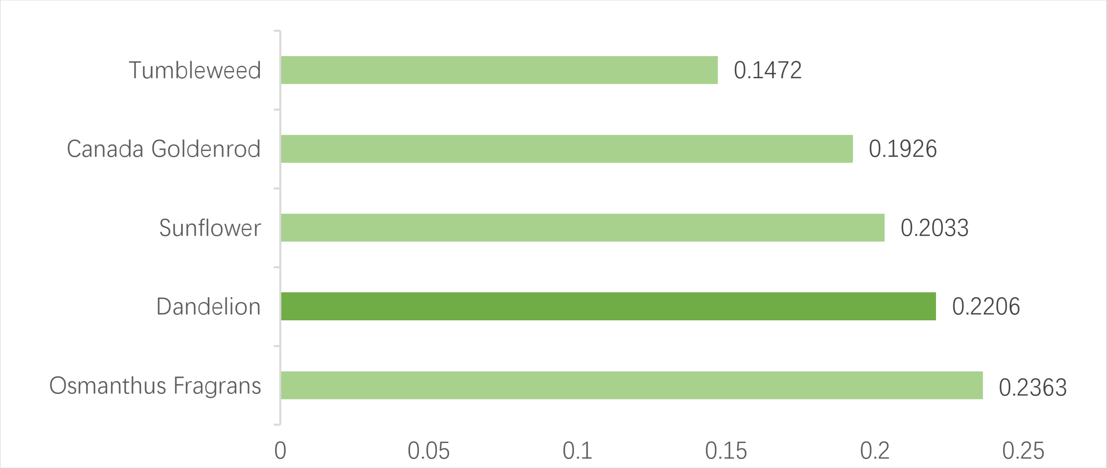

Figure 11. Final Scores (after normalization)

4.4 Model Results and Analysis

We chose five sets of data to build the model: two sets of non-invasive plant species (osmanthus fragrans and sunflower), two sets of invasive plant species (tumbleweed and Canada Goldenrod), and one set for dandelions. The final score of dandelions computed by topsis is 0.2206; tumbleweed and Canada goldenrod, which are two invasive plant species, scored 0.1472 and 0.1926 respectively; sunflower and osmanthus fragrans, which are two noninvasive plant species, scored 0.2033 and 0.2363 respectively. The results of the model showed us that dandelion was less invasive than one of the non-invasive vegetation-sunflower, suggesting that dandelion is not an invasive species. The result produced by the model is also consistent with the fact that plant has a rich history of medicinal and culinary use.

According to the weight calculated by entropy-method and result of cluster analysis, we found that the impact factors that determines an invasive plant species are the seeds per generation of the investigated species, the time required to produce a generation, the economic impact, the maximum and minimum temperatures, and the maximum and minimum humidity. Among these impact factors, the time it takes to reproduce one generation has the greatest influence on determining an invasive plant species, with a weight computed by the entropy method of 21.95%, the highest among the 7 impact factors. The factor with the second-largest influence is the maximum tolerance humidity, which carries a weight of 16.23%. Following this, the third factor is the minimum tolerance humidity, with a weight of 15.89%. Subsequently, the economic influence holds a weight of 15.68%. Next in line is the highest tolerance temperature, with a weight of 15.58%. The fifth factor is the lowest tolerance temperature, with a weight of 8.06%. The impact factor with the smallest influence on determining invasive plant species is the number of seeds per generation, which has a weight of 6.6%.

5 Model Evaluation

5.1 Strengths

a) The figures are very clear and easy to understand, and the scatter plot can visually show the distribution of seeds spread each month.

b) The density curve clearly shows the distribution density of the seeds under different temperature and humidity conditions.

c) Both of the model that we used in problem 2 have simple steps and they are easy to understand.

d) We used the clustering model to help determine the variables that we used in the TOPSIS model and it made the variables more convincing.

e) The use of entropy method that help us determine the weight made our determination being more objective instead of giving a weight to each variable by ourselves subjectively.

5.2 Weaknesses

a) The curves can only predict a single reproduction and seeds spread, providing limited information for long-term research.

b) In the process of establishing the model, the influence of wind speed on seed spread was ignored, and only the influence of temperature and humidity was noticed and focused on.

c) The algorithm of clustering method has a flaw that outliers and noisy data can affect the center point easily and cause an uncertainty.

d) The TOPSIS model can only process the quantitative index so we lose the accuracy on the qualitative index part.

Reference

[1] Wu, J. (2019). Probability model of seed dispersal distance in plants based on wind speed random variable. *Mathematics in Practice and Understanding*, 49(22), 129-144.

[2] Brogan, C. (2022). Engineers uncover secret 'thinking' behind dandelions' seed dispersal. Imperial College London.

[3] Seed Dispersal: Deciding when to move. (2023, January 31). Dandelion seeds respond to wet weather by closing their plumes, which reduces dispersal when wind conditions are poor. eLife, 12, e85477. https://doi.org/10.7554/eLife.85477

[4] Knops, J. M. H., & Lenda, M. (n.d.). Goldenrod honey: misinformation is causing a biological invasion of this Canadian weed. Retrieved from The Conversation website: https://theconversation.com/goldenrod-honey-misinformation-is-causing-a-biological-invasion-of-this-canadian-weed-152255.

[5] May 20, E. O. O. E., & Now, 2011 L. T. E. P. S. D. (2011, May 20). It may be High Noon for tumbleweed. Retrieved November 13, 2023, from www.hcn.org website: https://www.hcn.org/wotr/it-may-be-high-noon-for-tumbleweed.

[6] McClosky, J. W. (2014). Invasive Plant Early Detection in the San Francisco Bay Area (U.S. National Park Service). Retrieved November 14, 2023, from Nps.gov website: https://home.nps.gov/articles/invasive-plant-early-detection.htm

[7] Tu, W., Xiong, Q., Qiu, X., & Zhang, Y. (2021). Dynamics of invasive alien plant species in China under climate change scenarios. *Ecological Indicators*, 129, 107919.

Appendix

```
#%% Simulate the density curve of dandelions for the first spread
data = []
for i in np.arange(0, 20.1, 0.001):
 x = i
 A = (Q/(x*(sigema_lny)*np.sqrt(2*math.pi)))
 B = np.exp(-(np.log(((F*x)/H)/y_mean_g)/(np.sqrt(2))*sigema_lny)**2)
 dQ = A*B
 data.append(dQ)
 print(dQ)
 #Ey = np.exp(miu_lny+0.5*sigema_lny**2)
 #nameda_m = (Ey*H)/F

#%%
data = pd.DataFrame(data)
data = data.rename(columns={0:'Number of seeds'})
data.index = np.arange(0, 20.1, 0.001)
#%%
#plot the graph
plt.figure(figsize=(12,10), dpi = 600)
plt.plot(data['Number of seeds'],lw = 4,label = 'Dandelion')
plt.xlabel('Distance (m)',fontsize = 25)
plt.ylabel('Number of seeds (seeds)',fontsize = 25)
plt.title('The number of seeds changes with distance ',fontsize = 25)
plt.xticks(fontsize = 25)
plt.yticks(fontsize = 25)
plt.legend(loc = 'best',fontsize = 25)
#plt.grid(True)

# Set the horizontal axis to display all integers
ax = plt.gca()
ax.xaxis.set_major_locator(MaxNLocator(integer=True, prune='both'))
plt.savefig("./Distribution of dandelion.svg",dpi=600)
numbers = [dQ]
max_number = max(numbers)
print(max_number)
plt.show()

2
#%%Predict the distribution of first spread
data_dis_1 = []
```

```
 for q_1 in tqdm(range(Q)):
        x_1,y_1 = random_point_in_circle(nameda_m)
        data_dis_1.append([x_1,y_1])
        #Store the x and y coordinates in two separate lists
        x_coords_1 = [point[0] for point in data_dis_1]
        y_coords_1 = [point[1] for point in data_dis_1]
        #Create Scatter Chart
        # plt.figure(figsize=(12,12), dpi = 100)
        # plt.scatter(x_coords_1, y_coords_1, label='Scatter Plot', color='red', lw =
9,marker='*')
        #Add titles and labels
        # plt.title('Scatter Plot of (x, y) Coordinates')
        # plt.xlabel('X-axis')
        # plt.ylabel('Y-axis')
 3
Z = unnamed ./ repmat(sum(unnamed.*unnamed) .^ 0.5, n, 1);
disp(Z)

for i = 1:n
       for j = 1:m
              Z(i,j) = [unnamed(i,j) - min(unnamed(:,j))] / [max(unnamed(:,j)) -
min(unnamed(:,j))];
       end
end
disp(Z)
weight = Entropy_Method(Z);
disp(weight)
```