Team Number:	apmcm2110716
Problem Chosen:	C

2021 APMCM summary sheet

Construction of Ecological Conservation and Assessment of Its Impact on Environment

Abstract

For problem 1, in order to build a model analyzing how the Saihanba restoration have influenced environmental conditions, the paper first uses **Analytic Hierarchy Process** (AHP) to identify potential representatives of ecological environment, then uses **SVM** model to optimize, and finally employs **Entropy Method** to construct an environmental assessment index to quantify impact brought by Saihanba. Although the meteorological data set is relatively small, the model gives an optimistic **impact index 0.15**.

For problem 2, the paper analyzes the correlation between the historical frequency of sandstorms in Beijing and the restoration of Saihanba by comparing sandstorm conditions before and after the construction of the Tree Farm. **SVR model** is trained and identifies two significant with **accuracy level of 47.3%**.

For problem 3 and 4, the paper extends the research scope of the previous two problems and analyze the ecological condition and carbon emission efficiency of provinces in China and Mongolia and employs **Grey Relational Analysis** to detect the relationship between carbon emission and potential factors, which can best reflect the relative change in the afforestation process in the past decades. Finally, recommendations on the location and scale of new reservation spots are proposed based on the model.

According to the above analysis, the paper puts forward the main conclusion that reservation areas have a significant, positive influence on the surrounding ecological environment and are also conducive to improving the environmental condition of distant locations. Therefore, it is strongly suggested that afforestation, restoring green land should be considered the highest priority.

Keywords: SVM AHP Entropy Weight Method Grey Relational Analysis

Contents

I. Problem Background and Restatement

In order to indicate the origin of problems, the following background is worth mentioning:

1.1 Problem Background

Afforestation is a process where new forests are planted across land without trees, thereby extending the land's former short production cycles to much longer ones of forested lands, which is one of the most effective means of tackling climate change and protecting the environment. Under the guidance of the concept that lucid waters and lush mountains are invaluable assets, in Weichang Manchu and Mongol Autonomous County of China emerges a remarkable achievement of afforestation in the world's ecological civilization history Saihanba Tree Farm. As the largest man-made plantation globally, the construction of Saihanba Tree Farm had transformed 1.12 million acres of degraded land into a green paradise and eventually benefited surrounding cities such as Beijing from the perspective of sandstorm blocking, clean water supply, and carbon sequestering. As the pioneer of ecology restoration, the Saihanba Tree Farm has launched three new strategies on afforestation, natural improvement of artificial forestry, and near-naturalization cultivation of natural forestry since the holding of the 18^{th} National Congress of China's Communist Party, aiming to reinforce its ecological impacts further. Hence, understanding and quantifying the impact of the restoration of Saihanba on the local environment and ecology protection over time is of great importance and can guide further sustainable development around the country and even the entire globe.

1.2 Problem Restatement

Based on the above background, the team collected historical data on Saihanba's forest coverage, Weichang climate, Beijing sandstorm record, etc., and the following questions are answered:

1) Establish a mathematical model and collect relevant data to quantitatively analyze how the environmental condition has changed before and after the construction of Saihanbai Tree Farm and evaluate impact of the restoration.

2) Establish a mathematical model and collect relevant data to quantitatively analyze how the afforestation of Saihanba is related to the sandstorm resisting in Beijing and

evaluate impact of the restoration.

3) Extend the research scope of previous two problems, build a mathematical model to identify geographic locations in need of additional ecological reservation and corresponding scale, and evaluate its influence on achieving carbon neutral target.

4) Apply the model in problem three to quantitatively analyze another country's environment condition and identify locations and scale of potential ecological area to be built.

II. Problem Analysis

2.1 Problem One Analysis

Problem one requires the establishment of a mathematical model to quantify the local environmental change prior to and posterior to the construction of Saihanba Tree Farm through relevant data on both forestry farms and surrounding regions. To achieve the goal and erect an impact evaluation system, the team adapted a three-step method with the use of Analytic Hierarchy Process (AHP), Support Vector Machine (SVM), and Entropy Method, etc.

1) Step 1: To begin with, it is necessary to select and extract the principal indicators that quantitatively reflect the conditions of Saihanba Tree Farm and the ecological environment of surrounding regions. The team decides to utilize the AHP method to fulfill the needs.

2) Step 2: Secondly, a predictive model is established to predict the impact of Saihanba's daily-changing parameters on essential component indicators of the local ecological system. Considering the relatively small sample size and limited meteorological data, SVM is employed to optimize model performance.

3) Step 3: Finally, with the prediction model depicting the impact of Saihanba on essential ecological indicators separately, a concluding ecological assessment index is necessary to facilitate the overall evaluation of Saihanba's impact on local ecology and future prevalent application of the model. Because of the many advantages and common use of the entropy method in constructing an environmental assessment index, the paper utilizes the entropy method to establish an assessment index and score the performance of Saihanba.

2.2 Problem Two Analysis

Problem two requires the establishment of a mathematical model to quantify the sandstorm condition change in Beijing prior to and posterior to the construction of Saihanba Tree Farm through relevant data on forestry farm, distant regions, and interaction data between two locations (i.e. distance). To achieve the goal, historical and contemporary meteorological data are extracted for Saihanba and Beijing respectively. SVR is employed to achieve better performance on prediction due to its advantage in relatively low-volume and multiple distribution types of data. Through the estimated correlation coefficients, several significant features for the target variable are identified.

2.3 Problem Three Analysis

Problem three requires the establishment of a mathematical model to quantify the current carbon neutrality performance of different geographical locations and identify the ones demanding additional reservation. To achieve the goal, the team measures carbon neutrality performance of each province in China respectively and adapts a three-step method with the use of Grey Relational Analysis Model.

1) Step 1: Firstly, identify key factors of provincial carbon emission and use Grey Relational Analysis Model to calculate the correlation coefficient between each influencing factor and carbon emission.

2) Step 2: Secondly, apply the latest statistics such as population, GDP, and energy structure of each province to the above model to construct the indicator Carbon Emission (CE) representing the emission level of each province.

3) Step 3: Thirdly, select a benchmark to represent the carbon neutrality performance and compare the benchmark with each province's current ability for absorbing and sequestering carbon dioxide. Based on all the previous steps, the team proposes suggestions on the location and scale of building ecological areas.

2.4 Problem Four Analysis

Problem four requires the application of the relational analysis models established in problem three based on a foreign country's environmental data corresponding to the indicators determined in problem three. Therefore, the team uses a similar approach to problem 3 to identify locations and scale of ecological reservations to be built.

III. Assumptions

Historical data on forest, climate, and biodiversity are hard to collect, the following assumptions are necessary to clarify our research goal and quantify the uncertainty in the model and solution process.

1) It is assumed that the data published by Hebei Province and Beijing Meteorological Bureau are authoritative.

2) There exists a linear relationship between spatiotemporal variation of sandstorm, forest coverage, and environment factors.

3) It is assumed that sandstorm frequency is linearly correlated to indicators of local forest coverage ratio, wind speed, humidity, and precipitation.

4) It is assumed that total carbon emission is related to population, GDP, weight of secondary industry, urbanization ratio, and energy structure.

5). Ecological balance is measured by a variable visibility that indicates devastating effects of human pollution that destroys ecological balance.

* All the above assumptions will be further explained in the problem analysis section.

IV. Notations for Functions

See below Table 1 on page 5.

V. Model and Solution for Problem One

5.1 Component Analysis of Ecological Indicator System

5.1.1 *Analysis of Indicators Based on AHP*

Attributing to the many advantages of **Analytic Hierarchy Process(AHP)** application in quantifying complicated decision problems composed of numbers of interactive variables, the team decides to utilize AHP as the fundamental model to evaluate the performance of Saihanba on its surrounding ecological environment.

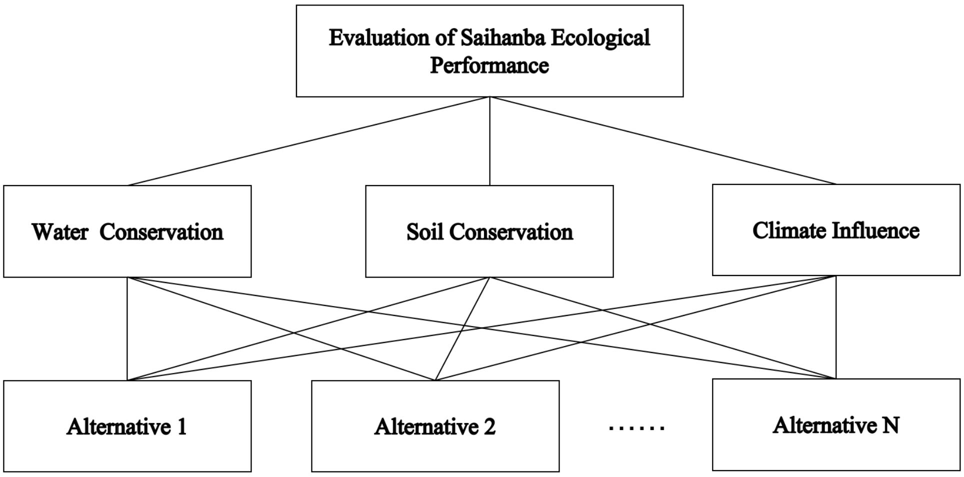
1) Construction of Recursive Hierarchical Model

To begin with, it is necessary to define the goal and identify decision criteria as well as alternatives in a traditional AHP method. In this case, based on the announcement

Table 1 Variables and their Descriptions

Symbol	Description
CE	Carbon Emission, a constructed indicator for carbon emission level
E/A	Emission divided by absorption
FAO	Forest cover ratio
A	A $n \times n$ matrix formed by all criteria variables assigned in both columns and rows
a_{ij}	The ratio between i^{th} variable in the row and j^{th} variable in the column
λ_{max}	The maximum eigenvalue of any matrix A
$\overrightarrow{v}_{max}$	The eigenvector corresponding to the maximum eigenvalue of matrix
CR	Consistency ratio to evaluate the AHP result
CI	Consistency index to evaluate the AHP result
RI	Random consistency index to evaluate the AHP result
n	The number of rows or columns of a matrix A
x_{ij}	The weight of j^{th} alternative variable under i^{th} criteria variable
w_{ij}	The weighted weight of j^{th} alternative variable under i^{th} criteria variable (standardized)
I_{ij}	The i^{th} observation of j^{th} alternative indicator
I_{max}	The maximum value of the j^{th} alternative indicator
W_{ij}	The weight of the i^{th} observation I_{ij} on the j^{th} indicator set
e_j	The information entropy of j^{th} indicator I_j
v_j	The variation factor of j^{th} indicator I_j

published by Chinese Environmental Protection Bureau on the factors damaging the ecological environment, the team considers the evaluation of Saihanba's performance as the decision goal and concludes three main criteria with over 32 alternative variable types along with accumulative report reading and other secondary source research: water conservation, soil conservation, and climate influence, shown in the graph 1 below.



Graph 2 Recursive Hierarchical Model for Problem One

Notice: because the number of potential alternative variables under each criteria variable selected through research is relative high, around 30 in total, the detailed alternative variables are not listed here. After the correlation determination and variable filtering through support vector machine in the later section of 5.4, the team will present a more clear and comprehensive AHP structure table containing much more certain alternative indicators.

2) Construction of pairwise comparison matrix

Secondly, with the determined three criteria and expert opinions on differentiated criteria factors absorbed from previous research, the team applies the consistent matrix method proposed by Thomas L. Saaty and constructs a 3×3 pairwise comparison matrix:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

with the property of

$$a_{ij} = \frac{1}{a_{ij}}$$

By computing quantitative assessment indexes according to Saaty's reference table in the appendix, we get

$$A = \begin{bmatrix} 1 & 0.5 & 1 \\ 2 & 1 & 4 \\ 0.5 & 0.25 & 1 \end{bmatrix}$$

3) Consistency Test

After performing the AHP modeling and normalization of vector in SPSS, the team obtained the normalized eigenvector targeting the maximum eigenvalue λ_{max} of matrix A defined in step 2:

$$\overrightarrow{v}_{max} = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} = \begin{bmatrix} 0.286 \\ 0.571 \\ 0.143 \end{bmatrix}$$

To confirm the priority weight of criteria variables (water conservation, soil conservation, and climate influence), the team conducts the consistency test through the calculation of consistency ratio CR based on the value of consistency index CI, random consistency index RI, and the following formulas:

$$CR = \frac{CI}{RI}$$

$$CI = \frac{\lambda - n}{n-1} = \frac{\lambda - 3}{3-1}$$

$$RI = \frac{CI_1 + CI_2 + \ldots + CI_n}{n} = \frac{CI_1 + CI_2 + CI_3}{3}$$

Implementing the calculation of data, the team obtained:

$$CR = \frac{CI}{RI} = \frac{0.000}{0.520} = 0.000 < 0.1$$

result indicates that the result has passed the consistency test and the team finally formulated a more clear AHP table containing criteria variable with confirmed priority weight and alternative variables leaving to be evaluated in the following section with the use of support vector machine and entropy method:

Based on the above selection and analysis of criteria variables through qualitative research and quantitative modeling, the team determines to extract and discover Saihanba's data from the three perspectives of water conservation, soil conservation, and climate influence in order to evaluate its performance reflected by the change of relevant statistics under each certain criteria.

5.1.2 *Alternative Indicator Determination*

After confirming the three main criteria attributes, the team subdivides each attribute into more detailed and specific alternative indicators based on qualitative research on each field. In the case of water conservation, the most commonly cited indicators

Table 2 Variables and their Descriptions

Criteria	Criteria Weight	Alternative Variables	Alternative Weight	Overall Weight
Water Conservation	28.6%	Soil Thickness	x_{11}	w_{11}
		Forest Biomass	x_{12}	w_{12}
		River Distance	x_{13}	w_{13}
		Forest Stand Origin	x_{14}	w_{14}
Soil Conservation	57.1%	Soil Condition	x_{21}	w_{21}
		Forestry Type	x_{22}	w_{22}
		Landscape Coverage	x_{23}	w_{23}
		...	x_{24}	w_{24}
Climate Influence	14.3%	Precipitation	x_{31}	w_{31}
		Wind Direction	x_{32}	w_{32}
		Cloud Amount	x_{33}	w_{33}
		...	x_{34}	w_{34}

are soil thickness, forest biomass, river distance, and forest stand origin. Considering the log-linear relationship between forest coverage and other features, the factor forest coverage is ignored in the following analysis. On the other hand, the team selects over 15 alternative indicators for the criteria of soil conservation and climate influence respectively based on research: soil condition, forest type and etc. for soil conservation, and precipitation, wind direction, etc. for climate influence. Due to the relatively large number of alternative indicators, the team decides to conduct SVM on the effect of each indicator on ecological features first, and later utilizes the entropy weight method to determine the indicator weight, constructing a final widely applicable evaluation index of any natural reservation.

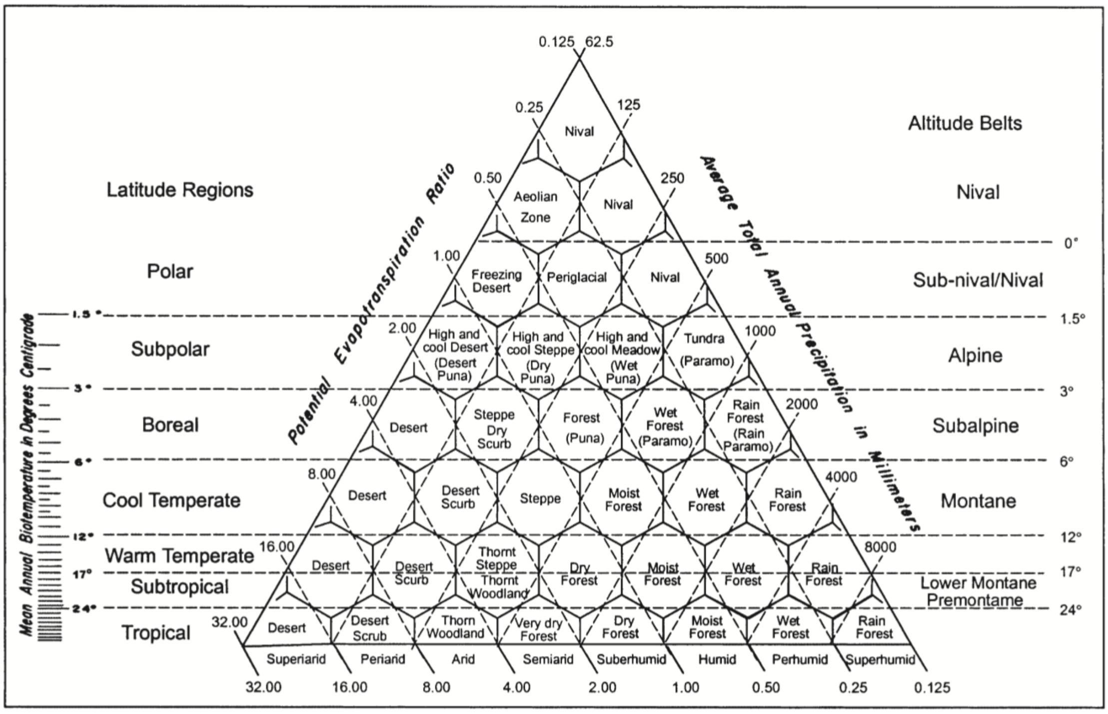
5.1.3 *Forest Coverage Indicator*

Forest coverage, defined as the percentage of soil covered by green forest, is an important indicator of ecological balance. However, according to a theory proposed by American botanist Leslie Holdridge in 1947, the forest coverage can be estimated by annual average temperature and precipitation.

Specifically,

$$PER = \frac{\frac{\sum t}{365} \times 58.93}{P}$$

where PER stands for potential evaporation rate, t stands for the mean daily temperature($^\circ C$), P stands for the annual precipitation(mm), and $\frac{\sum t}{365} \times 58.93 = PET$ where PET is the potential evaporation. Therefore, PER has a log-linear relationship with temperature and precipitation. Therefore the significance of forest coverage can be estimated by these two factors. Cited from a paper(Y.Pan) by a team from Beijing Normal University, Holdridge's life zone is classified as below:



Graph 1 Holdridge's life zone classification and diagram

5.1.4 *Sandstorm indicator*

5.2 Regional Ecological Data Acquisition

Meterological data for problem one is acquired through National Centers for Environmental Information (`https://www.ncei.noaa.gov/access/search/data-search`). Data for weather station of Weichang, Hebei in year 1977 and 2011 are downloaded to compare the impact of restoration of Saihanba on the environment.

5.3 Prediction of Principal Indicator Based on SVM Regressor

The second layer of the algorithm is achieved through **Support Vector Regression(SVR)**. SVR works by finding a hyperplane to fit the data with defined flexibility that allows a certain amount of error to be acceptable in the model, whereas all support vectors are data points that defines the position and margin of the hyperplane. In contrast to ordinary least squares regression, SVR minimizes the *l2*-norm of the coefficient vector instead of the squared error. The error term is handled in the constraint, i.e.the set flexibility. The user sets a value to be the maximum error such that absolute errors are all less than or equal to this margin. The model aims to minimize $\frac{1}{2}|w|^2$ with the constraint $|y_i - w_i x_i| \le \epsilon$.

When the data pattern is not linear, SVM finds a way to determine the hyperplane in a higher dimension where the transformation to the higher dimension is determined by a kernel function. The four kernel functions employed by the team are linear, polynomial, radial basis function(rbf), and sigmoid kernel functions, which are the four of the most common kernel functions that fits the data pattern best.

Through data preprocessing and feature selection two features are selected as the target variables of resisting sandstorms: daily average wind speed and precipitation. Through training and testing data, it is found that the SVR with kernel rbf has the best accuracy 50.4% for predicting sandstorm in year 1977, the SVR with kernel linear has the best accuracy 28.4% for predicting sandstorm in year 2011, the SVR with kernel rbf has the best accuracy 57.9% for predicting ecological balance in year 1977, and the SVR with kernel rbf has the best accuracy 26.7% in year 2011. Overall we selected 8 indicators for soil conservation and 9 indicators for environment influence for further analysis.

5.4 Construction of Ecological Assessment Index

5.4.1 *Determining Indicator Weight Based on Entropy Weight Method*

Considering the effectiveness and benefits of entropy weight method application in specific alternative variable weight determination, the team decides to employ the method construct an overall ecological assessment index for any natural reservation with six steps based on the AHP model established and improved through support vector machining.

1) Data Standardization

To confirm the weight of each alternative variable in a decision model based on AHP method, the team is required to determine the n indicators under each criteria variable and m samples which accumulate to build up a solid variable. To achieve the goal, the team conducts support vector machining to reduce the number of indicators n under the soil conservation criteria from 32 to 5 with the use of approximately 6200 samples of data.

Because of the dimension and unit difference across indicators, the team further processes the sample data by conducting the data standardization with the use of extreme value method according to the category of indicator, shown in the table below:

Depending on the category of each indicator I_{ij}, the team standardizes the data from each sample and restores them into a new csv file according to the following formulas on different conditions:

$$\text{If } I_{ij} \text{ is positive indicator, } I_{ij} = \frac{I_{ij} - I_{min}}{I_{max} - I_{min}}$$

$$\text{If } I_{ij} \text{ is positive indicator, } I_{ij} = \frac{I_{max} - I_{ij}}{I_{max} - I_{min}}$$

Table 3 Variables and their Descriptions

Criteria Indicator Category	Alternative Symbol	Alternative Indicator
Soil Conservation	T_i	Negative
	SD_i	Negative
	DP_i	Positive
	SLP_i	Positive
	$VISIB_i$	Negative
	$GUST_i$	Negative
	MAX_i	Positive
	MIN_i	Negative
Environment Influence	T_i	Negative
	SD_i	Positive
	DP_i	Negative
	SLP_i	Negative
	$WDSP_i$	Negative
	$GUST_i$	Negative
	MAX_i	Positive
	MIN_i	Positive
	$PRCP_i$	Negative

2) Construction of Weight Matrix

To obtain the value of information entropy of each indicator I_j, the team operates on individual indicator observation weight W_{ij} first and construct a weight matrix W of the entire dataset:

$$\mathrm{W} = \left[W_{ij}\right]_{m\times n} = \left[W_{ij} = \frac{I_{ij}}{\sum_{i=1}^{n} I_{ij}}\right]_{m\times n}$$

3) Determining the parameter for Each Indicator I_j

With the data on the weight of each observation which composed of the entire dataset for different indicators, the team was able to determine the information entropy e_j of each indicator I_j:

$$e_j = -k \times \sum_{i=1}^{n} W_{ij} \times \ln W_{ij}$$

with the formula of k as:

$$k = \frac{1}{\ln{(n)}}$$

and hence its variation factor v_j:

$$v_j = 1 - e_j$$

4) Construction of Weight of each Alternative Indicator under Same Criteria

To achieve the goal to assess the weight of each indicator using entropy weight method, the team utilizes the variation factor v_j of the indicator I_j obtained from information entropy e_j to evaluate the importance of this indicator based on the principle that the importance coefficient is positively correlated with the value of variation factor v_j:

$$W_j = \frac{v_j}{\sum_{j=1}^{m} v_j}$$

5) Calculation of Alternative Weight

After implementing the entropy weight method in SPSS, the team obtained a more comprehensive hierarchical table of evaluation decision problem:

6) Calculation of Overall Weight

With the statistics on both criteria weight and alternative weight, the team finalizes the hierarchical table with an additional column of overall weight:

Table 4 Variables and their Descriptions

Criteria	Criteria Weight	Alternative Variables	Alternative Weight	Overall Weight
Soil Conservation	57.1%	Temperature	22.76%	13%
		Snow Depth	0.00%	0.00%
		Dew Point	22.86%	13.05%
		Sea Level Pressure	0.01%	0.00571%
		Visibility	12.87%	7.35%
		Wind Gust	0.00%	0.00%
		Max Temperature	14.67%	8.38%
		Min Temperature	26.82%	15.30%
Climate Influence	14.3%	Temperature	2.66%	0.38%
		Snow Depth	0.00%	0.00%
		Dew Point	2.67%	0.38%
		Sea Level Pressure	0.00%	0.00%
		Wind Speed	4.51%	0.64%
		Wind Gust	0.00%	0.00%
		Max Temperature	1.71%	0.245%
		Min Temperature	3.13%	0.45%
		Precipitation	85.33%	12.20%

5.4.2 *Construction of Ecological Assessment System*

Inspired from ensemble algorithm, the team combined decision modeling of AHP, SVM, and entropy weight method to construct a hierarchical structure for the decision problem of Saihanba's ecological performance evaluation with well-estimated weight assigned on criteria variables and cleaned alternative variables. To further quantify and improve the evaluation system of Saihanba's ecological performance, the team determines to establish an ecological assessment index, taking the weight of confirmed criteria variables and alternative variables though previous modeling into account, with the hope to satisfy more prevalent usage needs. Hence, the ratios of Saihanba's performance in each confirmed indicator from the table in section 5.5.2 between each two years were extracted using Python, among which the average(A_h) was selected respectively to develop a weighted improvement ratio using the following formula:

$$I_{ecology} = W_1 \times I_1 + W_2 \times I_2 + ... + W_n \times I_n$$

5.4.3 *Evaluation of Saihanba's Ecological Performance*

After the performing of ratio calculation, the team collected the average ratios for all confirmed alternative indicators shown below:

Table 5 Variables and their Descriptions

Criteria	Alternative Indicator	Average Ratio Change	Indicator Weight
Soil Conservation	Temperature	−1.67%	13%
	Snow Depth	0.00%	0.00%
	Dew Point	−1.22%	13.05%
	Sea Level Pressure	0.12%	0.00571%
	Visibility	18.05%	7.35%
	Wind Gust	0.00%	0.00%
	Max Temperature	0.11%	8.38%
	Min Temperature	−3.45%	15.30%
Environment Influence	Temperature	−1.67%	0.38%
	Snow Depth	0.00%	0.00%
	Dew Point	−1.22%	0.38%
	Sea Level Pressure	0.12%	0.00%
	Wind Speed	36.02%	0.64%
	Wind Gust	0.00%	0.00%
	Max Temperature	0.11%	0.245%
	Min Temperature	−3.45%	0.45%
	Precipitation	89.16%	12.20%

Computing the data from the above table, the team obtains the value of evaluation index for Saihanba's ecological performance:

$$I_{ecology} = W_1 \times I_1 + W_2 \times I_2 + ... + W_n \times I_n = 4.3\% + 11.08\% = 0.15$$

With the range of index as 1, the value of ecological impact index for Saihanba as 0.15 is relatively low; however, considering the fact that water conservation, which weighs 0.29 among the three criteria indicators, is not discussed in this essay, 0.15

can be viewed as good performance within the range of index without water conservation consideration as 0.71. In addition, in terms of distant ecological impacts, it is satisfactory Saihanba as an immobile natural reservation to recover from heavy and frequent sandstorms, reaching the impact level over surrounding and local environment by approximately 15%.

VI. Model and Solution for Problem Two

6.1 Introduction of Sandstorm Formation Mechanism and Methodology

Beijing, China has long suffered from sandstorm caused by strong winds carrying dust from drought-hit Mongolia and other parts of northwest China in the history. Large-scale deforestation, strong wind, and drying weather are usually considered a factor of sandstorm. When a sandstorm hit, the level of PM2.5 can largely exceed the World Health Organization's average daily concentrations standard.

Since sandstorms only happen a few times each year, it is extremely hard to find daily indicator on storm. However, extensive research has been conducted and models have been constructed regarding the relationship between sandstorm and meteorological factors. The team cites the model constructed in Spatiotemporal variation of sandstorm and its response to forest restoration in Beijing-Tianjin sandstorm source area' to build a sandstorm indicator and using which to measure Saihanba's impact on sandstorms in Beijing.

6.2 Beijing Environmental Data Acquisition

Beijing climate data from 2005 to 2020 is collected from website Weather in the World (`https://rp5.ru/åŻťåœžéŢĞåŐĘåŔšåďŤæřŤ_`). The original data contains hourly information on the temperature, atmospheric pressure, humidity, wind direction, wind speed, cloud height, visibility, soil condition, precipitation, snow depth etc. of Weichang, Hebei where the forest farm locates and Beijing, the target site of sandstorm.

Through data preprocessing and feature selection the team computes the daily average for each indicators of the forest farm and target site respectively. The remaining eight features are input features of SVM: T, P_s, H, ff, T_l, T_h, T_{dew}, and P. The three target variables indicating sandstorms are: average wind speed at a height of 10-12m above ground during the first 10 minutes of observation,relative humidity at 2m above ground level, and relative humidity at 2m above ground level.

6.3 Prediction of Principal Indicator Based on SVM Regressor

After normalization that improves model performance, the model is defined in SVM regressor (SVR) with four possible kernel functions: linear, poly, rbf, and sigmoid. Support Vector Machine is a powerful supervised machine learning algorithm for classification and regression models. It works well with multiple types of distributions of data with few assumptions on the data. Considering the relative small size of dataset, the team decides to employ SVR given its efficiency.

Moreover, since performance under linear kernel function is acceptable, the team estimates the correlation coefficients of the eight features through weights the model assigned to them in prediction.

6.4 Results and sensitivity analysis

Through training and testing the SVR with rbf kernel function achieves a mean accuracy of 47.3% on the testing data without further parameter tuning. With parameter tuning the model would easily achieve a performance with over 50% accuracy. Given the limited size in data set, the team considers the result considerably inspiring.

By Table 5, there are 4 features that contributes the most to prediction: the atmospheric temperature at 2m above ground level, the atmospheric pressure at mean sea level, the relative humidity at 2m above ground level, and the average wind speed at a height of 10-12m above ground during the first 10 minutes of observation.

Table 6 Features of SVR model and their weights

T	P_s	H	ff	T_l	T_h	T_{dew}	P
-1.37	-0.19	-0.36	0.30	1.00	0.16	0.21	0.07

Of the four, the negative sign of temperature and positive sign of wind speed are in accordance with existing conclusions from papers that sandstorm is more likely to appear with lower temperature and higher wind speed. The model reveals some insights from data that has not been widely noticed: that pressure at sea level and precipitation has a relatively high importance in predicting sandstorms as well.

6.5 Discussion: Distance between Saihanba and Beijing

On the other hand, the team supposes that other factors such as the distance of Beijing from Saihanba are also of significance in determining if there is going to be sandstorm in Beijing. However, due to the limitation of algorithms, it is hard to be quantified in the model.

Common algorithms, both machine learning algorithms and data mining algorithms, are built upon the variation of features. Features such as distance does not have variation and thus would be considered insignificant by model because it cannot explain changes in target variables. However, it is obvious that the value of distance affects how the forest farm Saihanba protects Beijing from sandstorm. It is a limitation of the algorithm that it cannot account for such factors.

VII. Model and Solution for Problem Three

7.1 Introduction of Carbon neutrality Mechanism

At its address to the 75^{th} session of the UN General Assembly in September 2020, President Xi Jinping declared that China would 'aim to have CO_2 emissions peak before 2030 and achieve carbon neutrality before 2060.'

Carbon neutrality is a state of net-zero carbon dioxide emission, which can be achieved by balancing emissions of carbon dioxide with an equivalent amount of carbon mitigation or by eliminating emissions from society. Afforestation engineering programs that promote increased terrestrial carbon stocks are an important means to help gradually decrease atmospheric CO_2 emissions since a new forest ecosystem through afforestation that absorbs carbon dioxide from the atmosphere and fixes it in forest and soil, thereby reducing the concentration of carbon dioxide in the atmosphere.

7.2 Methodology

The team measures each province's demand on new ecological reservation based on the gap between carbon emission and current ability to absorb carbon. Through exploring the topic the team learned that China's studies on carbon emission has been lagging leading to limited data. Accordingly, the team decided to employ a **Grey Relational Analysis Model** that works well with little information and small sample size.

Grey relational analysis is built upon grey relational degree and determining the contribution measure of the main behavior between the system factors. Assume there are S subject and each subject has N indicators, define

$$x_i(\text{m, n})$$

as the n^{th} indicator of the m^{th} subject observed at time t. Then, the data matrix of indicator i, X_i is

$$\begin{bmatrix} \mathrm{x}_i(1,1) & \mathrm{x}_i(1,2) & \ldots & \mathrm{x}_i(1,\mathrm{N}) \\ \mathrm{x}_i(2,1) & \mathrm{x}_i(2,2) & \ldots & \mathrm{x}_i(2,\mathrm{N}) \\ \ldots & \ldots & \ldots & \ldots \\ \mathrm{x}_i(\mathrm{M},1) & \mathrm{x}_i(\mathrm{M},2) & \ldots & \mathrm{x}_i(\mathrm{M},\mathrm{N}) \end{bmatrix}$$

Initialize the data matrix by setting initializer d_i as

$$\frac{1}{x_i(1,1)}$$

Then, initialize data matrix d_iX_i to be

$$\begin{bmatrix} 1 & x_i(1,2)d_i & \ldots & x_i(1,N)d_i \\ x_i(2,1)d_i & x_i(2,2)d_i & \ldots & x_i(2,N)d_i \\ \ldots & \ldots & \ldots & \ldots \\ x_i(M,1)d_i & x_i(M,2)d_i & \ldots & x_i(M,N)d_i \end{bmatrix}$$

The grey relative coefficient y_i(m, n) between d_iX_i and d_1X_1 can be calculated by

$$\frac{D + p \times L}{|x_i(m,n)d_i - x_1(m,n)d_1| + (p \times L)}$$

where p is resolution and

$$\mathrm{L} = \min |\ \mathrm{x}_i(m,n)d_i - x_1(m,n)d_1|$$
$$\mathrm{U} = \max |\ \mathrm{x}_i(m,n)d_i - x_1(m,n)d_1|$$

Based on the assumption that population, urbanization, total production, weight of secondary industry, energy structure are the most essential factor of carbon emission. Due to the lack of data in carbon emission, a grey relative coefficient data matrix between total carbon emission and each of the factors developed by a team from Dalian University of Technology is cited below. Correlation coefficient can be calculated with $\mathrm{r}_{i,m} = \frac{\sum_{n=1}^{\infty} y_i(m,n)}{N}$.

Table 7 Grey Relational Coefficient

	Population	GDP	Urbanization Rate	Second Industry Weight	Energy Structure
Carbon Emission	0.786	0.806	0.729	0.770	0.774

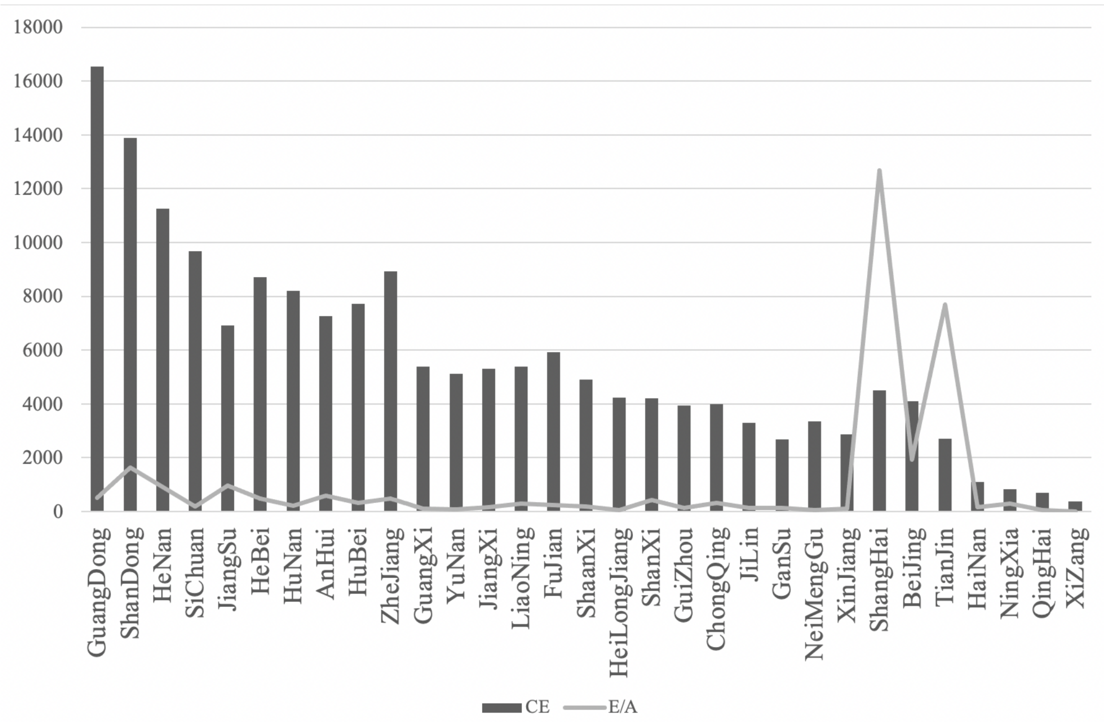
population in 10,000 and GDP in billion

7.3 Determining Target Province and Reservation Location

Applying statistics of each province to the above model, the team constructs an indicator Carbon Emission representing emission level of each province. As is shown in the following graph, Guangdong, Shandong, Henan, Sichuan, and Jiangsu Province rank top 5 in terms of carbon emission level. However, instead of simply comparing total emission, the team also employs

$$E/A = \frac{CE}{VAO \times Area}$$

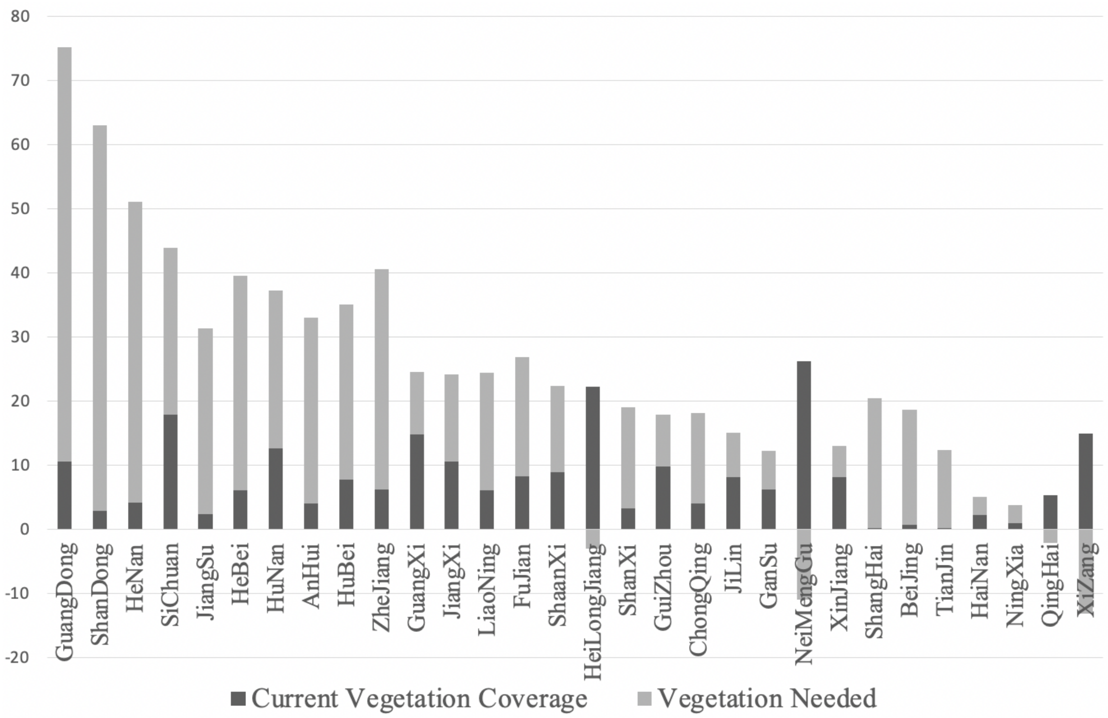
to measure the discrepancy between carbon emission and carbon absorption. Even though Beijing, Shanghai, and Tianjin have relative low carbon emission, which is large due to low weight on secondary industry and favourable energy structure, their E/A ratio is significantly higher than other provinces, indicating a huge gap between carbon emission and the ability to absorb, due to the high urbanization rate and lack of ecological area.



Graph 2 Each province's carbon emission level

From Graph 2, one can also observe that provinces such as Guangxi, Yunnan, Xizang have remarkably low E/A ratio indicating a good performance. Since it is hard to determine the threshold of achieving carbon neutrality, the team determines the amount of additional ecological area each province needs in order to have a performance in line with the performance of Yunnan Province.

The result is presented in Graph 3. For example, the forest area in Sichuan Province is 17.923 units while additional 26.044 units is needed to achieve self carbon neutrality.



Graph 3 Current forest coverage vs additional forest needed

VIII. Solution for Problem Four

8.1 Background and Country Selection

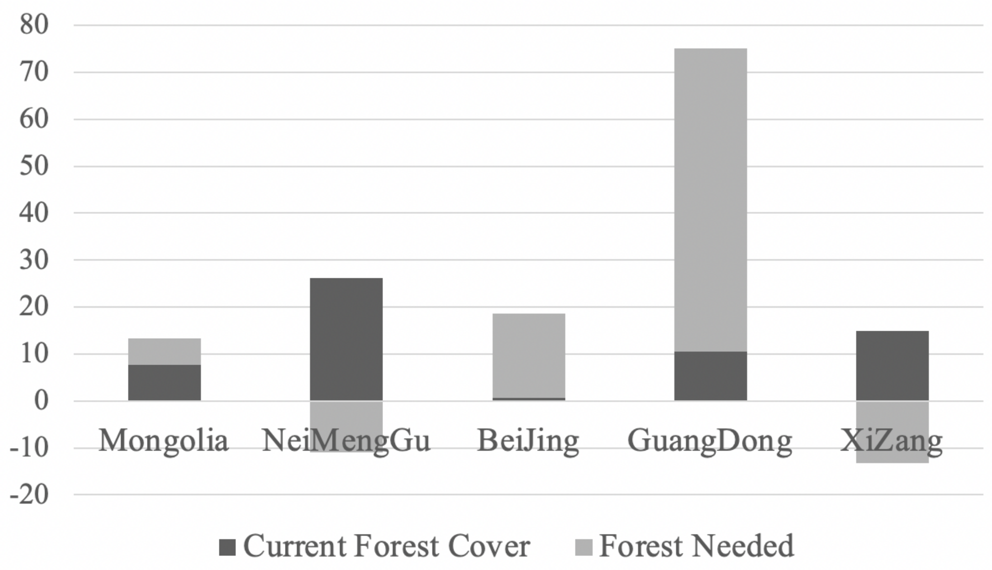
In mid March this year, the northern part of China experienced sandstorm with longest duration, strongest intensity, and widest impact in the past decade, and the sandstorm was mainly the result of the influence of a Mongolian cyclone and a cold front on the ground. Moreover, since the formation of sandstorm requires basic conditions including strong winds, dusty sources and unstable atmospheric structure, among which the sand and dust in the northwest region is mainly contributed by Mongolia. Therefore, the team decides to analyze ecological environment of Mongolia.

8.2 Analysis

Table 8 Compare statstics of different locations

Location	Mongolia	Neimenggu	Beijing	Guangdong	Xizang
Population	327.8	2534	2170.7	11346	337.15
GDP	13.14	1728.922	3031.998	9727.777	147.763
Urbanization	0.7	0.6271	0.865	0.707	0.3114
Second Industry Weight	0.3827	0.5542	0.227	0.4854	0.3464
Energy Structure	0.36	0.404	0.775	0.701	0.42
Forest Cover Ratio	0.079	0.221	0.435	0.5859	0.1214
Area	156.65	118.3	1.64	17.98	122.84

The team collects data from Statista (`https://www.statista.com/`) and uses the approach proposed in problem three to construct the CE indicator and E/A for Mongolia. Sample data is shown below, and it can be easily observed that Mongolia's forest coverage ratio is considerably lower than that of the other locations.



Graph 4 Current forest coverage vs additional forest needed

The result is presented in Graph 4 that the current forest area in Mongolia is 7.676 units while additional 5.715 units is needed to achieve self carbon neutrality (again using the E/A ratio of Yunan province as a benchmark). Based on the conclusion of problem one and problem two, afforestation may not only improve local ecological environment, but also have positive impact on other nearby provinces and countries.

IX. Summary and Potential Improvements

9.1 Conclusions to previous problems

- **9.1.1** ***Problem one***

  Due to the breadth and complexity of problem one, the team designed an ensemble algorithm that contains AHP, SVM, and entropy weight method to produce an overall Saihanba performance of 0.15 out of 1 in benefiting the surrounding environment. Considering the limited data for this process, the outcome is indicating a promising future for Saihanba.

- **9.1.2** ***Problem two***

  The team implements SVM in evaluating the effectiveness of Saihanba in protecting Beijing from sandstorm with a model of over 47% accuracy. The model complies common conclusions while reveals insights on previously unnoticed coefficients as well.

- **9.1.3** ***Problem three***

  In order to decide which provinces have the most imbalanced carbon emission that needs to be neutralized with a forest farm in problem three, the team employs a Grey relational analysis model to determine carbon emission of each province in China through several indicators including population, GDP, Urbanization Rate etc. The model best works with the restricted development of carbon emission studies in China.

- **9.1.4** ***Problem four***

  The team extends the model and conclusions in problem three to Mongolia given its frequency and large size of impact in suffering from sandstorm. Comparing its natural resources in forest and vegetation also produces supporting evidence of the suggestion.

9.2 Applications of the models in future plans for ecological reservations

According to all above results, the team suggests that the goal of carbon neutralization and resisting sandstorms should be considered with the highest priority in planning ecological reservations. Evidence from models support that building reservations are effective measures in protecting the environment and surrounding cities.

Moreover, in selecting locations for the reservation, modeling outcomes suggest that factors such as visibility and precipitation as well as commonly known temperature and wind speed are of significance. On the other hand, contracting results of models suggest to avoid base conclusions on a single model. The ensemble algorithm employed in problem one might be worthy in such circumstances.

9.3 Potential improvements

Along with the problem of algorithms with constant factors such as distance discussed in solutions to problem two, exploring through the project leads the team to realize the difficulty of applying algorithms in reality. Besides the limited volume in data, the algorithm does not perfectly solve for the desired quantities of the problems. Instead, they are often viewed as a reference for estimation. This leads the team into thinking the limitations of algorithm in reality.

Moreover, the team believes that problems can be solved with more sufficient data, such as data from ECMWF(`https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5`) which has more abundant and comprehensive data but are stored in the obscure grib file format. Due to the limited time, the team decided not to spend the majority of time working on data but looked more for information that supports modeling. Theories must be grounded upon solid data evidence. Therefore the team believes with better data input the models can produce more reasonable and insightful conclusions.

X. References

[1] Y.Qin et al., Spatiotemporal variation of sandstorm and its response to vegetation restoration in Beijing-Tianjin sandstorm source area[J], 10.3969/j.issn.1002-6819.2012.24.027

[2] P.Xiong et al., Grey correlation analysis of carbon emissions in East China, Place of Publication: Press, Year of publication, 10.19525/j.issn1008-407x.2021.01.005

[3] J.Liu et al.,Land-cover classification of China: integrated analysis of AVHRR imagery and geophysical data, Vol 24, No.12, 2485–2500

[4] L.Jiang et al., In situ soil moisture and temperature network in genhe watershed and saihanba area in China, 10.1016/j.dib.2020.105693

[5] B.Duan et al., Cause Analysis on Severe Dust Storm in Northern China on 15 March 2021, Vol.39: 541-553.

XI. Appendix

Table 9 Variables names and description for data in problem one

Symbol	Description
Meteorological data	
T	Average temperature
SD	Snow Depth
DP	Dew point
SLP	Sea Level Pressure
$WDSP$	Wind Speed
$VISIB$	Visibility
$GUST$	Wind Gust
MAX	Maximum temperature
MIN	Minimum temperature
$PRCP$	Precipitation

Table 10 Variables names and description for data in problem two

Symbol	Description
T_{2m}	Normalized Difference forest Index
P_s	Daily average precipitation
H	Relative humidity (%) at 2 meters above the earth's surface
ff	Air temperature at 2 metre height above earth's surface
T_l	Lowest temperature in the past 12 hours
T_h	Highest temperature in the past 12 hours
T_{dew}	Dew point temperature at 2m above the ground
P	Precipitation

Listing 1: The python Source code of one case of Data Processing

```
from google.colab import drive
drive.mount('/content/drive')
```

```
import pandas as pd
bj = pd.read_excel('/content/drive/MyDrive/2021 apmcm/Q2_bj.xls')
bj['preci'] = bj['preci'].fillna(0)
bj['date'] = bj['real_time'].str[:10]
bj['date'] =
    bj['date'].str[6:]+'/'+bj['date'].str[3:5]+'/'+bj['date'].str[:2]
bj = bj.sort_values(by = 'date')
bj.reset_index(level=None, drop=True, inplace=True, col_level=0,
    col_fill="")
bj = bj.groupby(by=['date']).mean()
bj.sort_index().round(3)
bj.reset_index(inplace=True)
bj.reset_index(inplace=True)
shb = pd.read_excel('/content/drive/MyDrive/2021 apmcm/Q2_shb.xls')
shb['preci'] = shb['preci'].fillna(0)
shb['date'] = shb['real_time'].str[:10]
shb['date'] =
    shb['date'].str[6:]+'/'+shb['date'].str[3:5]+'/'+shb['date'].str[:2]
shb.sort_values(by = 'date')
shb.reset_index(level=None, drop=True, inplace=True, col_level=0,
    col_fill="")
shb = shb.groupby(by=['date']).mean()
shb.sort_index().round(3)
shb.reset_index(inplace=True)
# check dates in one dataset that does not have records in another
bj[~bj.date.isin(shb.date)]
bj = bj.drop([708, 709, 808, 1528, 1529, 1530])
bj.reset_index(level=None, drop=True, inplace=True, col_level=0,
    col_fill="")
compare = bj.eq(shb)
compare['date'].value_counts() # all values are true showing that date
    records of two dataframes are in accordance
y = bj[['avg_windspeed', 'preci', 'humid']]
# dealing with missing values
nan_values = shb.isna()
nan_columns = nan_values.any()
columns_with_nan = shb.columns[nan_columns].tolist()
```

```
print(columns_with_nan)
shb['Po'].isna().sum() / len(shb['Po']) = # 0.435
shb['Ps'].isna().sum() / len(shb['Po']) = # 0.0096
shb['Pa'].isna().sum() / len(shb['Po']) # = 0.485
shb['max_windspeed'].isna().sum() / len(shb['Po']) # = 1.0
shb['max_speedbt'].isna().sum() / len(shb['Po']) # = 0.853
shb['low_tem'].isna().sum() / len(shb['Po']) # = 0.0130
shb['hi_tem'].isna().sum() / len(shb['Po']) # = 0.0127
shb['preci_t'].isna().sum() / len(shb['Po']) # = 0.656
shb['soil_temp'].isna().sum() / len(shb['Po']) # = 1.0
shb['snow_depth'].isna().sum() / len(shb['Po']) # = 1.0
shb['soil_condna'].isna().sum() / len(shb['soil_condna']) # = 1.0
# drop columns that have over 10% of missing values
# impute mean values for the remaining columns with missing values
x = shb.drop(columns=['date', 'Po', 'Pa', 'max_windspeed',
    'max_speedbt', 'preci_t', 'soil_temp', 'snow_depth', 'soil_condna'])
x.loc[:,"hi_tem"] =
    data.loc[:,"hi_tem"].fillna(data.loc[:,"hi_tem"].mean())
x.loc[:,"low_tem"] =
    data.loc[:,"low_tem"].fillna(data.loc[:,"low_tem"].mean())
x.loc[:,"Ps"] = data.loc[:,"Ps"].fillna(data.loc[:,"Ps"].mean())
```

Listing 2: The python Source code for SVM

```
# just attach code for one model here since the code for training and
    testing SVM models are structurally identical
from sklearn.model_selection import train_test_split
# set features: x and target variables: y
# split testing and training datatest_siz decides the ratio of splitted
    dataand random_state decides the mode of random splitting
Xtrain, Xtest, Ytrain, Ytest =
    train_test_split(x,y,test_size=0.2,random_state=42)
for i in [Xtrain, Xtest, Ytrain, Ytest]:
    i.index = range(i.shape[0])
# normalization
from sklearn.preprocessing import StandardScaler
cate = Xtrain.columns[Xtrain.dtypes == "object"].tolist()
col = Xtrain.columns.tolist()
```

```python
for i in cate:
    col.remove(i)
ss = StandardScaler()
ss = ss.fit(Xtrain.loc[:,col])
Xtrain.loc[:,col] = ss.transform(Xtrain.loc[:,col])
Xtest.loc[:,col] = ss.transform(Xtest.loc[:,col])
from sklearn.feature_selection import VarianceThreshold
selector = VarianceThreshold()
X_Var0 = selector.fit_transform(x)
X_Var0.shape # all eight features have enough variation in themselves
    to predict changes in target variables
from sklearn.svm import SVR
from sklearn.metrics import r2_score
Ytrain = pd.DataFrame(Ytrain)
Ytest = pd.DataFrame(Ytest)
Ytrain = Ytrain.iloc[:,0].ravel()
Ytest = Ytest.iloc[:,0].ravel()
kernels = ['linear','poly','rbf','sigmoid']
for kernel in kernels:
    svr = SVR(kernel=kernel).fit(Xtrain,Ytrain)
    Ypredict = svr.predict(Xtest)
    print(kernel+"kernel SVM has an accuracy of", svr.score(Xtest,
        Ytest))
    print(kernel+"kernel SVM has R_squared", r2_score(Ytest, Ypredict ))
# rbf kernel SVM returns the best result
svr = SVR(kernel='linear').fit(Xtrain,Ytrain)
Ypredict = svr.predict(Xtest)
svr.coef_
```

Listing 3: The python Source code for Model of Problem Three

```python
import pandas as pd
carbon = pd.read_excel('carbon.xlsx')
carbon['Carbon_Emis_Score'] = 0.768 * carbon['Population'] + 0.806 *
    carbon['GDP'] + 0.729 * carbon['Urbanization'] + 0.77 *
    carbon['Secondary_industry'] + 0.774 * carbon['Energy_Efficiency']
carbon.sort_values(by = ['Carbon_Emis_Score'], ascending=False)
output = carbon[['Province','Carbon_Emis_Score']]
```

```
import matplotlib.pyplot as plt
plt.figure(figsize=(15,4))
output.plot(kind='bar',x='Province',y='Carbon_Emis_Score', color = '00')
carbon['Green'] = carbon['Coverage'] * carbon['Area']
carbon['ratio'] = carbon['Carbon_Emis_Score'] / carbon['Green']
carbon['target'] = 220.294322
carbon['need'] =
    carbon['Carbon_Emis_Score']/carbon['target']-carbon['Green']
carbon['result'] = carbon['need']/carbon['Green']
```