The Need for Bees: Colony Prediction and Analysis based on ARIMA and Entropy Weight -TOPSIS

Shizhou You [1], Sipeng Yan [2], Zhenghan Xu [2], Jiahao Hu [3]

[1] Nanjing Foreign Language School, Nanjing City, 210000, China

[2] Hohhot No.2 High School, Hohhot city, 010000, Canada

[3] The stony brook school, NY,11790, USA

Abstract. In this paper, firstly, the ARIMA time-series differential autoregression algorithm was used to establish the change model of the number of bee colonies, and then the entropy weight method TOPSIS and Pearson correlation matrix were fused to analyze the index sensitivity. Finally, the prediction model of the number of bee colonies was established based on BP dynamic neural network, to complete the prediction and index analysis of bee colonies. Finally, we will discuss the findings from mathematical modeling of the blog summary writing, which show that the colony's optimal temperature, climate, feed supply, and other environmental conditions have a significant impact on the rate of soft production.

Keywords: ARIMA; Entropy Weight -TOPSIS; Pearson Correlation Coefficient; BP Dynamic Neural Network; Sensitivity Analysis.

1. Introduction

Honeybees are beneficial insects to humans as well as an important class of economic insects. Honeybees not only serve humans with honey and other bee products, but they also pollinate agricultural crops. Approximately 90% of fruit trees rely on bee pollination and are closely connected to humans. The biological pollination of honeybees has greatly improved the quality and yield of crops and maintained ecological balance. In 2004, the US published a review of honeybee genome sequences stating that "without honeybees, entire ecosystems would collapse". Bee pollination is an evolutionary and interdependent process that cannot be substituted by any human means to improve productivity, making bees a crucial component of contemporary agriculture. Bees require pollen and nectar for their production and reproduction, and plants generate these substances. Despite the fact that bees have numerous ecological advantages, honeybee populations consistently experience population reductions. There are a variety of factors contributing to the decline in bee populations, including shortened lifespans brought on by the intense summer workload (overactivity), bacterial and fungal infections, dietary factors affecting longevity, excessive use of pesticides that kill bees, and environmental effects. Whatever the situation, the main variables that contribute to bee mortality are the environment, people, and self-activity.

Whatever the reason for the decline in the bee population, it usually brings some trouble. To quote physicist Albert Einstein, if bees disappeared from the surface of the earth, humans would have only four years of life left because food crops would be unpollinated and bees are responsible for pollinating one-third of our crops, such as trees, berries, vegetables, and even alfalfa, but not other crops such as corn, tubers and root vegetables (like potatoes and carrots) or leafy vegetables. In nature, about 5% of plant species are heavily dependent on these organisms. In fact, honeybees have a major impact on food production, and if they were to become extinct, their impact on humans would be immense.

2. Model I: Colony Number Change Model based on ARIMA Temporal Differential Autoregression

We chose the total population of China's Zhejiang province as a colony area, the population with time change itself is a time series. We can know the population data by MATLAB mapping, and

based on the known data to forecast the future data, here is chosen by ARIMA difference autoregressive model for the population trend prediction.

Multi-step prediction of the number of bee colonies in different time series is carried out by ARIMA. Firstly, the time series model of the number of bee colonies is simply considered ARIMA $(p, 1, \mathrm{d})$, where $W_t = Y_t - Y_{t-1}$, then:

$$W_t = \phi_1 W_{t-1} + \phi_2 W_{t-2} + \cdots + \phi_{t-p} W_{t-p} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_{t-q} e_{t-q} \; t = \pm 1, \pm 2, \ldots$$

That is:

$$Y_t - Y_{t-1} = \phi_1 (Y_{t-1} - Y_{t-2}) + \phi_2 (Y_{t-2} - Y_{t-3}) + \cdots + \phi_{t-p} \left(Y_{t-p} - Y_{t-p-1}\right) + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_{t-q} e_{t-q} \; t = \pm 1, \pm 2, \ldots$$

We can get:

$$Y_t = (1 + \phi_1) Y_{t-1} + (\phi_2 - \phi_1) Y_{t-2} + (\phi_2 - \phi_3) Y_{t-3} + \cdots + \left(\phi_p - \phi_{p-1}\right) Y_{t-p} - \phi_p Y_{t-p-1} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_{t-q} e_{t-q} \; t = \pm 1, \pm 2, \ldots$$

The specific steps of the whole algorithm are as follows:

Plot the number of bee colonies data, showing the known state of the number of bee colonies, and observe whether it is a stationary time series. For non-stationary time series, D-order difference operation is first performed, which is converted into a stationary time series.

3. Model II: Sensitivity Analysis of Fusion Entropy Weight Method TOPSIS and Pearson Correlation Matrix

In order to conduct a sensitivity analysis of the model and determine which factors have the greatest impact, we need to consider two aspects, namely the degree of importance and the corresponding correlation. We take into account factors such as lifespans, egg-laying rates, fertilized/unfertilized egg ratios, weather conditions, the availability of food for the colony, disease factors, temperature conditions, colony status, queen status, and the number of feeding bees.

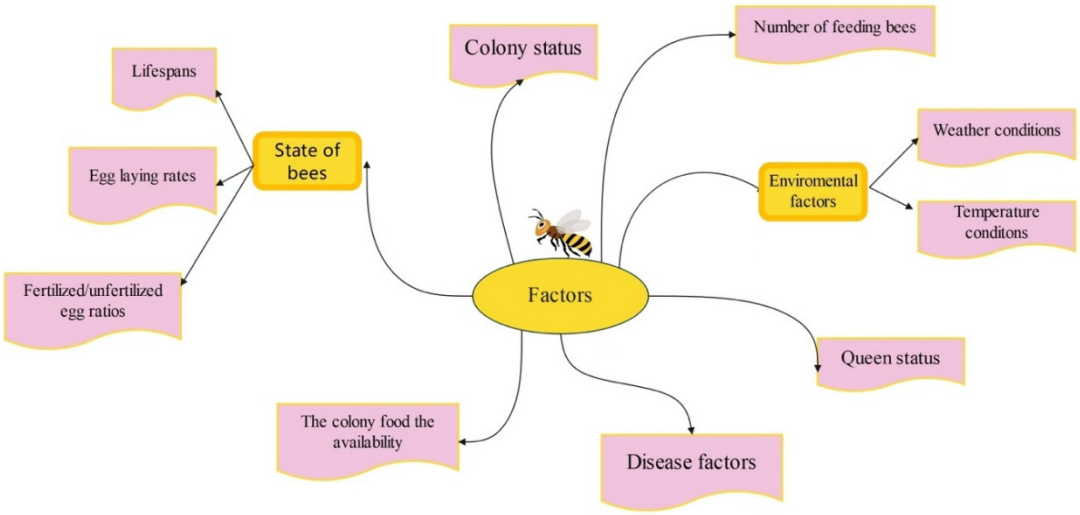


Figure 1. Evaluation index of the bees

As for the 10 indicators, some of them are strongly correlated, so the dimensionality of the indicators is firstly considered to be reduced by thermal mapping and K-means clustering analysis.

Any two index data are denoted as X_i and Y_i respectively, and the covariance of X and Y is calculated, namely:

$$Cov(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1} = \frac{\sum_{i=1}^{n} (X_i - \frac{1}{n}\sum_{i=1}^{n} X_i)(Y_i - \frac{1}{n}\sum_{i=1}^{n} Y_i)}{n - 1}$$

Pearson's correlation coefficient can be further obtained by dividing the covariance by the standard difference of X and Y, i.e.

$$a_{XY} = \frac{Cov(X,Y)}{S_X S_Y} = \frac{Cov(X,Y)}{\sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{i=1}^{n} X_i\right)^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2}{n-1}}}$$

The final m×m (10×10) -dimensional Pearson correlation coefficient matrix A can be obtained by pairwise cross-correlation of all indexes, and the element a_{ij} represents the correlation between index i and index j, i.e

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mm} \end{pmatrix}$$

Based on this matrix, the thermal map can be made to complete qualitative analysis, and then further K-means clustering analysis can be carried out. On the one hand, the verification results can be tested, and on the other hand, the characteristics of bee colony influencing factors can be further analyzed. The specific algorithm process is as follows:

1. Randomly select K colony index sample data as the clustering center.
2. Calculate the distance between each colony index data sample and each cluster center, using Euclidean distance here.
3. Conduct regression of each colony index data sample to the nearest cluster center.
4. Calculate the mean value of the colony index sample of each class and use it as the new clustering center.
5. If the class center doesn't change or reaches a set number of iterations, the algorithm ends. Otherwise, go back to step 2.

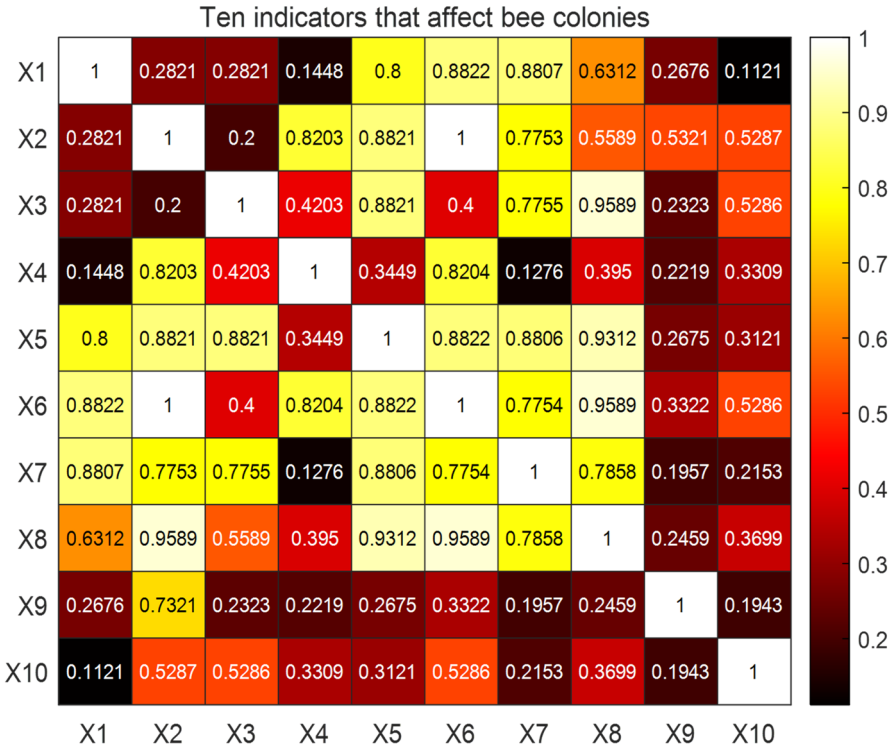


Figure 2. Correlative thermal map of ten indicators of bee colony

After the establishment of Pearson correlation matrix, the clustering results

In this way, the qualitative modeling of the thermal map and K-means clustering analysis is completed. Then, the five selected indicators need to be given quantitative weights, and the entropy weight method is used to assign weights. The specific algorithm steps are as follows:

Step 1: Index digitization and forward. For climate conditions in the colony, it is necessary to digitize, for example, good weather is quantified as 1, bad weather is quantified as 0, and $X_1 \ldots . X_{10}$ index is converted into a forward index, and the forward matrix $X = \{x_{ij}\}_{n\times 10}$; is obtained.

Step 2: Normalize the forward data. The standardized matrix $Z = \{z_{ij}\}_{n\times 10}$, including:

$$z_{ij} = x_{ij} / \sqrt{\sum_{i=1}^{n} x_{ij}^2}$$

Step 3: The normalized TOPSIS efficiency score was calculated. Is dimensionless and uniformly standardized, and the normalized efficiency score is performed. Define the maximum $Z^+ = (Z_1^+, Z_2^+, \cdots, Z_m^+) = (\max\{z_{11}, z_{21}, \cdots, z_{n1}\}, \cdots, \max\{z_{1m}, z_{2m}, \cdots, z_{mm}\})$. Define the minimum

$$Z^- = (Z_1^-, Z_2^-, \cdots, Z_m^-) = (\min\{z_{11}, z_{21}, \cdots, z_{n1}\}, \cdots, \min\{z_{1m}, z_{2m}, \cdots, z_{mm}\})$$

Define the $i(i = 1,2, \cdots, n)$ of evaluation objects and the maximum distance:

$$D_i^+ = \sqrt{\sum_{j=1}^{m} \left(Z_j^+ - z_{ij}\right)^2}$$

Define the $i(i = 1,2, \cdots, n)$ of evaluation objects and the minimum distance:

$$D_i^- = \sqrt{\sum_{j=1}^{m} \left(Z_j^- - z_{ij}\right)^2}$$

Then, we can calculate the normalized score of the $i(i = 1,2, \cdots, n)$ evaluation object:

$$S_i = \frac{D_i^-}{D_i^+ + D_i^-}$$

And the final S_i is the TOPSIS efficiency score of each corresponding bee colony.

Step 4: Calculate the ith a colony of the proportion of the first j indicators:

$$y_{ij} = \frac{x_{ij}'}{\sum_{i=1}^{m} x_{ij}'} \text{ (j=1,2,3,4,5)}$$

Step 5: Calculate the information entropy of the jth index:

$e_j = -K\sum_{i=1}^{m} y_{ij} \ln y_{ij}$, K is constant and $K = \frac{1}{\ln m}$

Step 6: Calculate the weight of the jth index:

$$w_j = \frac{1 - e_j}{\sum_j 1 - e_j}$$

Therefore, the entropy weight method model based on the thermal map and K-means clustering is established. Firstly, the correlation of indicators is defined by thermal map and K-means clustering, and then the index is weighted by the entropy weight method. Finally, qualitative and quantitative weights and correlations can be obtained as shown in Table 1 below.

Table 1. Influencing factors of bee colony

Index	Indicators	Weights	Correlation	Sensitivity
1	lifespans	0.91	0.99	0.95
2	egg-laying rates	0.94	0.98	0.96
3	fertilized/unfertilized egg ratios	0.95	0.99	0.97
4	weather conditions	0.92	0.89	0.91
5	colony food availability	0.81	0.86	0.84
6	disease factors	0.91	0.92	0.92
7	temperature conditions	0.85	0.81	0.83
8	colony status	0.71	0.82	0.76
9	queen status	0.86	0.87	0.87
10	number of feeding bees	0.81	0.91	0.86

Based on the above analysis, it can be seen that different factors have different effects on the number and size of bee colonies, and egg-laying rates have the greatest effect on this number.

4. Model III: Prediction Model of Beehive Number based on BP Dynamic Neural Network

The swarm can act to pollinate crops. For different areas where the colony number is different, the corresponding hives are also different. To predict how many hives are needed to support 20 acres of pollination (81,000 square miles) of land that benefits from crop pollination, we can use the dynamic BP neural network training to obtain the ideal solution and optimize the number of hives. For the bee colony data, the past data of the number of beehives and the corresponding land area are selected as the training set, and the regression model in the dynamic neural network is used to predict the situation of different beehives and areas, namely, the input sequence data of m beehives are denoted as $x_1 - x_m$, where x is a key-value pair, denoted as x_i=(u_i, w_i). u_i is the area of the ith case, where w_i is the number of beehives in the ith case, and the output is l predicted data denoted as $o_1 - o_l$, o_i=(u_i, w_i).

The neural network model can be further implemented by the following steps, input layer 1×m vector:

$$X = (x_1, x_2, \dots, x_i, \dots, x_{\mathrm{m}});$$

The hidden layer vector of n ($l = 1,2,3 \dots n$)

$$H^l = \left(h_1^l, h_2^l, \dots, h_j^l, \dots, h_{s_l}^l\right)(l = 2,3, \dots, L-1, j = 1,2, \dots, s_l)$$

Output vector of output layer:

$$O = (o_1, o_2, \dots, o_i, \dots, o_{\mathrm{l}})$$

Let w_{ij}^l be the connection weight between the ith neuron in layer $l-1$ and the jth neuron in layer l, b_j^l be the bias of the jth neuron in layer l, and net_j^l be the input of the JTH neuron in layer l, $f(\cdot)$ is the activation function, set the activation function to sigmoid, that is:

$$f(\cdot) = \frac{1}{1+e^x}$$

Thus, we get:

$$h_j^l = f\left(\mathrm{net}_j^l\right) = \frac{1}{1+e^{\mathrm{net}_j^l}}$$

$$net_j^l = \sum_{j=1}^{s_{l-1}} w_{ij}^l + b_j^l$$

Define the error function:

$$E = \frac{1}{m}\sum_{i=1}^{m} E(i)$$

where $E(i)$ is the training error of a single sample:

$$E(i) = \frac{1}{2}\sum_{k=1}^{2} (d_k(i) - y_k(i))^2$$

Therefore, the global error function is:

$$E = \frac{1}{2\times m}\sum_{i=1}^{m}\sum_{k=1}^{2} (d_k(i) - y_k(i))^2$$

Finally, the parameters are solved by gradient descent, and the final best binary classification effect can be obtained by adjusting the hyperparameters to continuously optimize the solving process. The updating process of weight and bias is as follows, where α is the learning rate and $\alpha \in (0,1)$:

$$w_{ij}^{l} = w_{ij}^{l} - \alpha \frac{\partial E}{\partial w_{ij}^{l}}$$

$$b_{j}^{l} = b_{j}^{l} - \alpha \frac{\partial E}{\partial b_{j}^{l}}$$

The overall training state is good, and optimization can be achieved through gradient descent and BP, and the corresponding optimal number of beehives can be found for different land areas.

After training, we only need to input a certain area into the program to get the corresponding optimal number of hives, so we can input 20 acres to get the number of hives. The following number of hives were required simultaneously for various land areas.

Table 2. The quantitative relationship between different hives and land area

Land area (acres)	Honeybee hives	Land area (acres)	Honeybee hives
1	5	15	79
2	11	20	110
4	21	30	178
6	33	50	297
10	53	100	600

In summary, to support 20 acres of pollinated land (81,000 square meters) for crops that benefit from pollination, 110 hives are required, with a dynamic range of 95 to 125 hives.

5. Conclusion

5.1 Strength

In this paper, the ARIMA time series model and the entropy weight TOPSIS evaluation model BP neural network prediction model have produced satisfactory solutions. At the same time, when small noise data is introduced, the model can also represent the results well, and the model has good generalization ability and strong robustness.

Among them, the ARIMA prediction model can effectively forecast small sample data while mitigating the disadvantage of the extensive computational of the operating models. TOPSIS and the entropy weight approach, when paired with K-means clustering analysis, can provide a more acceptable, effective, and thorough sensitivity analysis from two aspects of correlation and importance. All of this makes the bank risk assessment model more accurate, diversified, and effective, and the mutual qualitative and quantitative verification between the categorization and evaluation models increases the stability of the results.

Based on the traditional model of classic bold embedded in some innovative ideas, such as traditional TOPSIS combined with the entropy weight method, the BP neural network is introduced into the regularization vector optimization. In addition, we use the results of the forward difference to identify according to accurate characterization, setting the fuzzy tolerance to 2, according to the number of differences less than 0 to define whether the number of colonies negatively correlated with development. As a result, the overall model is accurate and efficient.

5.2 Weakness

ARIMA algorithm requires strong stationarity of time series. At the same time, the modification of hyperparameters for BP neural networks is more difficult, time-consuming, and labor-intensive. Additionally, because of the high level of uncertainty, it requires subjective prior intervention.

5.3 Possible Improvement and Promotion

Apart from the model discussed in this research, other approaches can be coupled to optimize and improve the prediction and analysis model of the bee colony. For more precise and efficient iterative optimization, the number of beehives can be determined using, such as the particle swarm

optimization method or genetic algorithm. Xgboost, fuzzy grey prediction, and other techniques can be used to further verify and optimize other evaluation and prediction models.

Furthermore, the model can be used to forecast and assess bee colonies, but it can also be expanded to forecast and assess the performance of other sectors. The categorization, assessment, and prediction of the chemical and energy industries, machine manufacturing, artificial intelligence, and other sectors can be further promoted by modifying the model's parameters and state.

References

[1] Baoyu, Tian, Nibal H, et al. Long-term exposure to antibiotics has caused accumulation of resistance determinants in the gut microbiota of honeybees. [J]. mBio, 2012.

[2] Period, T. F. "Experimental Wing Damage Affects Foraging Effort and Foraging Distance in Honeybees." (2011).

[3] Yan X, Han R C. Diagnostic technologies of common pathogens of honeybees in China[J]. Chinese Bulletin of Entomology, 2008, 45(3):483-488.

[4] Xiang-Jun L I. Time Series Models for Portfolio[J]. Journal of Wenzhou Teachers College, 2003.

[5] Yang Z B, Zhou R X, Zhang Q, et al. A Portfolio Optimization Model Based on Information Entropy and Fuzzy Time Series[C]// 2013 Sixth International Conference on Business Intelligence and Financial Engineering. 0.

[6] Meng X. Discount-Weighted Bayesian Model Averaging for Portfolio Decisions in Matrix Variate Dynamic Linear Models.

[7] Liu X T. Study on Data Normalization in BP Neural Network[J]. Mechanical Engineering & Automation, 2010.

[8] Huang Y, Meng S, Hwang S W, et al. Neural Network for Classification of Chinese Zither Panel Wood via Near-infrared Spectroscopy[J]. 2020(1).

[9] Li, Q, and Z. Xu. "Parameter Identification Method Research Based on the BP Neural Network and Space Search." 2014 International Conference on Advances in Materials Science and Information Technologies in Industry (AMSITI 2014) 0.