A Study on the Growth Characteristics of Green Algae Blooms in the Yellow Sea using Categorical Analysis

Yingli Tang, Shijun He
College of Information
Shanghai Ocean University
Shanghai, China
heshijun6@163.com

Haiyang He
Information Science and Engineering
East China University of Science and Technology
Shanghai, China

Peimin He
Institute of Marine Science
Shanghai Ocean University
Shanghai, China

Hairong Qi
Electrical Engineering and Computer Science
University of Tennessee
Knoxville, TN 37996
hqi@utk.edu

***Abstract*—Green algae (Enteromorpha prolifera) bloom, also known as the green tide, is becoming a new type of marine ecological disaster. More and more studies have been conducted to investigate the cause of green tide as well as its growth and drifting characteristics. However, most existing works either rely on single factor analysis or support vector machine-based multi-factor analysis that do not have an appropriate mechanism to handle categorical data. In this paper, categorical data analysis is adopted to study the effect of multiple environmental factors to the green algae evolution. Among the five factors we studied, temperature, wind force, and wave height have numerical values but weather condition and wind direction are nominal variables. We focus on the study of the correlation between these environmental factors and the size of distribution area of the green tide. Logistic regression models are built using remote sensing data from the Southern Yellow Sea acquired in 2013. Comprehensive analysis and evaluation are carried out, showing the main factors that affect the size of the bloom to be temperature and wave height, as well as their interaction. We also obtain the influence level of each main factor such that accurate prediction and early warnings of green tide can be made possible to alleviate the damage of the bloom.**

Keywords—categorical analysis; green algae blooms; environmental monitoring; multi-factor analysis; logistic regression model

I. INTRODUCTION

The world's largest green algae (Enteromorpha prolifera) bloom, covering an area of approximately 600 km^2 along the coastal area of Qingdao between May and July 2008, just weeks before the sailing competition of the 29th Olympic Games, has raised great concern [1-2]. At its peak offshore, the bloom covered 1,200 km^2 and affected 40,000 km^2 [1]. Several studies [3-4] have been conducted to investigate the cause and growth characteristics. So far, most studies have been limited to the single factor analysis method [5-7], which consider only one single impact factor of the green tide and cannot reflect the comprehensive effect of multiple factors. In addition, existing studies do not have an effective mechanism to handle categorical data.

Different from the traditional single factor analysis which is based on numerical data analysis, the categorical data analysis technique [8] is specifically efficient in dealing with data items that cannot be properly quantified or scaled to ensure the validity of the results more effectively. What's more, it gives full consideration to the correlation between multiple predictor variables, providing more rich information to the phenomenon studied. Categorical data analysis has been widely used in social science, behavior science, biomedical and many other fields [9-10]. This paper adopts the categorical data analysis technique to study the correlation between the distribution area of the green tide and five climate factors including temperature, weather condition, wind direction, wind force, and wave height. The analysis shows that temperature and wave height are the most important factors that influence the growth pattern of the green tide.

The rest of the paper is organized as follows: Section II describes the categorical analysis technique we adopted, including the definition of categorical data and the model construction process. Section III discusses experiments conducted, data sets used, as well as results obtained. Section IV concludes the paper.

II. CATEGORICAL ANALYSIS

A. Categorical Data

A *categorical variable* has measurement scale consisting of a set of categories. For example, choice of weather condition may use categories "sunny", "cloudy" and "rainy". Categorical variables have two main types of measurement scales. Many categorical scales have a natural order, e.g., response to a medical treatment (excellent, good, fair, poor). These are referred to as *ordinal scale*. Categorical variables with unordered scales are called *nominal variables*, such as weather condition mentioned above. Methods designed for nominal variables give the same results regardless the order of categories. Methods designed for ordinal variables utilize the category order, and results of ordinal analyses would change if the categories are listed in different order [8].

978-0-933957-43-5 ©2015 MTS

There are mainly three models used in categorical data analysis for trend analysis, namely, the linear probability model, the logistic regression model, and the probit model, among which, the logistic regression model is one of the most frequently used [11]. Logistic regression is originally designed to model binary response variables. In this paper, we extend the model to be able to handle multiple categories. According to whether the response variable is ordered, the model is divided into ordinal logistic model and nominal logistic model. Considering attributes of green tide, the paper adopts nominal logistic model.

Multiple categories response data can be analyzed by using logit model. Assuming Y is the response variable and the number of categories for Y is J. Let $\{\pi_1,...,\pi_J\}$ denote the response probabilities, satisfying $\sum_j \pi_j = 1$, it specifies the probability for each possible way the n observations can fall in the J categories. The logit models of the nominal response variable are to match each category with a baseline category. When the baseline is the last category (J), the baseline-category logit for variable x is:

$$\log(\frac{\pi_j}{\pi_J}) = \alpha_j + \beta_j x, j = 1,...,J-1 \quad (1)$$

Equation (1) determines equations for all other pairs of categories. The multicategory logit model has another expression of the response probability:

$$\pi_j = \frac{exp(\alpha_j + \beta_j x)}{\sum_h exp(\alpha_h + \beta_h x)}, j = 1,...,J \quad (2)$$

B. Model Selection

The stepwise variable selection algorithm [8] is adopted in this paper, as in ordinary regression, this algorithm can add or remove variables from a model in a stepwise manner. Forward selection adds variables sequentially until further additions do not improve the fitting. Backward elimination begins with a complex model and sequentially removes terms if such removal does not degrade the fitting. The process stops when any further removal leads to a significantly poorer fit. The selection process becomes more challenging as the number of variables increases because of the rapid increase in possible effects and interactions. We will experiment with the Akaike Information Criterion (AIC) which judges a model by how close its fitted values tend to be to the true expected values, as summarized by a certain expected distance between the two. The optimal model is the one that tends to have its fitted values closest to the true outcome probabilities, that is, the one that minimizes Eq. (3).

AIC = -2(log likelihood – number of parameters in models) (3)

Generally speaking, the log likelihood term measures the correlation between fitted model and true data, and the number of parameters aims to remove uncorrelated parameters or variables. In other words, lower AIC indicates better model fitting. In exploratory study, AIC can indicate which variable affects more on the model.

III. Experiment

A. Data Description

In this paper, we use the remote sensing data of Enteromorpha prolifera at the Southern Yellow Sea in 2013 [12-14] as our test data. According to [15], the factors that affect the evolution of green tide include temperature (T), weather condition (W), wave height (WH), wind force (WF), and wind direction (WD). The green tide distribution area (DA) is the response variable, which is divided into 5 categories, labeled as 1, 2, 3, 4, and 5, according to the size of the area. In this paper, the fifth category is chosen as the baseline category in the logistic regression. The explanatory variables are the five factors that affect the distribution area, among which W takes on three nominal values (“sunny”, “cloudy”, “rainy”) and WD takes on eight nominal values (E, W, S, N, NE, NW, WS, SE), and others are numerical data.

The data set is extracted from the weather report of meteorological observatory [16]. Table I shows a subset of the testing dataset, where the temperature value used takes the average temperature of that day.

TABLE I. Subset of the Testing Dataset

DA	*W*	*WD*	*WF (level)*	*WH(m)*	*T(□)*
1	sunny	W	4	1.2	17.5
1	sunny	W	4	1.2	18.5
1	cloudy	W	4	1.2	18.5
1	rainy	W	5	1.5	17
2	rainy	E	7	3.5	19.5
1	rainy	SE	4	1.2	21.5
1	rainy	SE	4	1.2	22.5
2	cloudy	SW	4	1.2	20.5
2	rainy	SE	4	1.2	18.5
1	cloudy	SE	4	1.2	18.5
5	cloudy	NE	4	1.2	21
3	cloudy	SE	4	1.2	23.5
3	cloudy	SE	4	1.2	24
3	cloudy	SE	4	1.2	24
3	cloudy	SE	4	1.2	24.5
3	rainy	E	7	3	22
2	rainy	NE	8	5	20.5
2	rainy	SE	4	1.2	26
4	rainy	S	4	1.2	25
4	rainy	SW	5	1.5	24.5
4	cloudy	W	4	1.2	27.5
1	cloudy	SW	4	1.2	28
1	rainy	SE	4	1.2	27

TABLE II. ENCODING OF WD, AND W

Class	*Value*	*Design Variables*
WD	NW	1000000
	N	0100000
	NE	0010000
	W	0001000
	E	0000100
	WS	0000010
	S	0000001
	SE	0000000
W	sunny	10
	rainy	01
	cloudy	00

In the logistic regression model, variables W and WD have nominal scales and need to be converted to quantitative data for further analysis. There are usually three coding schemes for converting nominal variables, including dummy coding, effects coding, and contrast coding. In the experiments, we choose the simplest of all, i.e., dummy coding. In dummy coding, a code of J-1 variables or bits is used where J is the number of categories that the nominal variable has. Usually one reference category is (randomly) identified and assigned the value of 0 for each code variable. The other categories are assigned a value of 1 for their specified code variable. Only one category can be assigned 1 for a specific code variable. Using this coding scheme, the eight categories of WD and three categories of W are converted using a code of 7 bits and 2 bits, respectively, as shown in Table II.

B. *Construction and Analysis of the Logistic Regression Model*

In practice, if we have no prior knowledge about the effects, it is not reasonable to add all variables into a model. The stepwise variable selection algorithm can be used to find the best model with the most effective variables. Specifically, WF and WH have a strong correlation (0.89115). So it is nearly redundant to use them both. In this study, we focus on T, W, WD, and WH.

We use SAS to construct the logistic model. In order to analyze the influence of various factors on the distribution of the green algae area, we need to verify whether the model's predicted values have high consistency with the corresponding values (i.e., goodness of fit). The commonly used index to indicate goodness of fit includes, for example, Pearson χ^2, deviance (D), likelihood-ratio (LR), and Hosmer-Lemeshow (HL) [17]. Pearson χ^2 and deviance are only applicable to the situation where the number of data belonging to each class is small. HL is usually used for logistic regression models that have many independent variables or contain continuous variables. Considering the nature of the dataset given, we choose to use the likelihood ratio index as the indicator of the model's goodness of fit.

The best model is obtained finally by using the stepwise variable selection algorithm. The explanatory variables, estimate of response parameters and likelihood-ratio test of the model are shown in Table III, where only the factors that generate a p-value (Pr > ChiSq) that is less than 0.2 are shown. That is, we set the alpha level to be 0.2, which is the threshold value that we measure p-values against. Any p-values less than 0.2 would mean that we need to reject the null hypothesis that there is no relationship between the corresponding predictor (or factor) and the response. "Intercept 1" through "Intercept 4" show the estimated ordered logits for the adjacent levels of the response variable, i.e., {5} versus {4, 3, 2, 1}; {5, 4} versus {3, 2, 1}; {5, 4, 3} versus {2, 1}; and {5, 4, 3, 2} versus {1}, respectively. As we can observe from Table III, the main factors affect the distribution area of the green tide are T and WH, as well as their interaction.

TABLE III. ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES

Parameter	**Estimate**	**Pr > ChiSq**
Intercept 1	-27.1865	0.1701
Intercept 2	-33.6930	0.1106
Intercept 3	-12.4514	0.5418
Intercept 4	-14.0791	0.4906
WH	20.4738	0.1845
WH	24.0425	0.1401
WH	7.5718	0.6308
WH	9.1486	0.5635
T	1.3027	0.1157
T	1.4938	0.0961
T	0.5245	0.5362
T	0.6423	0.4489
T*WH	-0.9121	0.1539
T*WH	-1.0521	0.1270
T*WH	-0.2906	0.6545
T*WH	-0.3827	0.5575

Although the optimal model is selected using the stepwise variable selection algorithm, the verification of validity is also needed in order to confirm the logistic regression model is appropriate. We illustrate the Wald, score tests, and likelihood-ratio by testing the null hypothesis (i.e., all coefficients equal to zero), the result is shown as Table IV.

TABLE IV. TESTING GLOBAL NULL HYPOTHESIS

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	19.4107	12	0.0791
Score	21.3684	12	0.0452
Wald	14.6458	12	0.2614

The above test results show that the regression model is effective and can be used to predict the response variable. That is to say, the reflective information of the model is also effective.

In order to contrast the influential level of various factors, the type 3 analysis of effects was adopted in this paper. Table V shows the test result of the hypothesis: $H_0 : \beta_j = 0$. It can be observed from Table V, T (P-value is 0.1240), WH (P-value is 0.138) and T*WH (P-value is 0.1114) interaction effect will affect DA. Again, an alpha value of 0.2 is used. The T*WH interaction affects the area distribution of the green algae the most. Since WF and WH have a strong correlation, that is to say, the T*WF interaction affects the DA the most. The second most influential factor is T and the last is WH. It confirms that WF is the main force of the evolution the green tide in the ocean and the sustained strong wind becomes the main external force of its drifting [6]. As we can observe from the prediction model, the coefficient of T is positive, this is also in line with the results obtained from [18] that T and the outbreak of the green tide has obvious relevance.

TABLE V. TYPE 3 ANALYSIS OF EFFECTS

Effect	T	WH	T*WH
Pr > ChiSq	0.1240	0.1388	0.1114

IV. CONCLUSIONS

This paper built up a predictive model of the distribution area of the Enteromorpha prolifera based on the logistic regression model according to the 2013 Enteromorpha prolifera remote sensing data in the Southern Yellow Sea. We studied five factors, T, WP, WD, WF and WH, that might have an influence on the evolution of the green tide. Through backward elimination selection algorithm, the model's factors that have no significant influence on Enteromorpha prolifera distribution were gradually eliminated and the main factors were obtained, including temperature, wave height, as well as the interaction between the two. We then obtained the correlation details (influence level) between the main climate factors and the migration of the green tide. With the learned correlation, the distribution of the green tide can be better predicted such that some preventive strategies can be taken to reduce the damage. At the same time, through the model test, it is obvious that this method is suitable for exploring the complex correlation between multiple response variables and predictor variables, making it a suitable tool for studying other marine phenomena.

REFERENCES

[1] D. Y. Liu, J. K. Keesing, Q. G. Xing, "World's largest macroalgal bloom casused by expansion of seaweed aquaculture in China," Marine Pollution Bulletin, 58(6):888-895, 2009.

[2] C. M. Hu , D. Q.Li, C. S.Chen, J. Z. Ge, "On the recurrent Ulva prolifera blooms in the Yellow Sea and East China Sea," Journal of Geophysical Research, 115, C:05017,2010.

[3] S. Song, F. Wang, C. L. Li, S. Qin, M. J. Zhou, et al., "Emerging challenges: Massive green algae blooms in the Yellow Sea," Nature Precedings, hdl:10101/npre.2008.2266.1.

[4] F. L. Qiao, G. S. Wang, X. G. Lu, D. J. Dai, "Drift characteristics of green macroalgae in the Yellow Sea in 2008 and 2010," Chinese Science Bulletin, 56(21):2236-2242, July 2011.

[5] D. P. Li, Y. Q. Yang, H. Y. Dong, "Cause Analysis and Synoptic Characteristics for Outbreaks of Entermorpha Around Qingdao in 2008", Periodical of Ocean University of China, 39 (6):1165-1170, 2009.

[6] L. Yi , S. P. Zhang,Y. Q. Yin, "Influence of Environmental Hydro-Meteorological Conditions to Entermorpha prolifera Blooms in Yellow Sea,2009," Periodical of Ocean University of China (Natural Science), 40(10):015-023,October 2010.

[7] S. P. Zhang, Y. C. Liu, G. Q. Zhang, L. Guan, "Analysis on the Hydro-Moteorological Conditions from Remote Sensing Data for the 2008 Algal Blooming in the Yellow Sea," Periodical of Ocean University of China, 39 (5):870-876, 2009.

[8] A. Agresti, "An Introduction to Categorical Data Analyssis," 2rd ed., New York: John Wiley and Sons,Inc., pp.68-73, 2007.

[9] S. Johnson, "Socio-economic Factors And Stage At Presentation Of Head And Neck Cancer Patients In Ottawa," Canada: A logistic Regression Analysis. 2010.

[10] H. John , "Goddeeris. Economic Outcomes In Young Adulthood For Extremely Low Birth Weight Survivors," Pediatrics. 126(5):1102-1108, 2010, .

[11] S. H. Wang, "Statistical analysis method and it's applications for the multi-response categorical data", Beijing University Of Technology, 2012

[12] X. F. Gu, X. F. Chen, Q. Yin, "Stereoscopic Remote Sensing Used in Monitoring Entermorpha Prolifra Disaster in Chinese Yellow Sea," Spectroscopy and Spectral Analysis, 31(6):1627-1632, 2011.

[13] J. Gower, C. Hu, G. Borstad, "Ocean color satellites show extensive lines of floating Sargassum in the Gulf of Mexico," IEEE Transactions on Geoscience and Remote Sensing, 44(12): 3619-3625, 2006.

[14] C. Hu, "A novel ocean color index to detect floating algae in the global oceans," Remote Sensing of Environment, 113(10): 2118-2129, 2009.

[15] H. Zhang, "Analysis of Outbreak Mechanism and Cntroll of the Green Tide in South Yellow Sea," Dalian: Danlian maritime University, 2013.

[16] http:/5/www.tianqihoubao.com/

[17] D. T. Bui, O. Lofman, I. Revhaug, "Landslide susceptibility analysis in the Hoa Binh province of Vietnam using statistical index and logistic regression," Natural Hazards, 59(3):1413-1444, 2011.

[18] S. L. Fan, M. Z. Fu, Y. Li, "Origin and development of Huanghai (Yellow) Sea green-tides in 2009 and 2010," Acta Oceanologica Sinica, 34 (6): 186-194, November 2012.