EARTH & ENVIRONMENTAL SCIENCES

Modeling of natural wetland methane emissions with machine learning

Hongjiu Zhang[1] (p930026154@mail.uic.edu.cn), Faculty of Science and Technology, BNU-HKBU United International College

ABSTRACT

Methane is very widely distributed in nature and is a major component of natural gas and biogas. Besides carbon dioxide, methane is the second largest anthropogenic contributor to global warming. With the rapid acceleration of global warming, it is necessary to effectively analyze the factors affecting methane emissions and accurately predict methane emissions. Therefore, this project develops a multiple linear regression model by applying the data detected at the site.

This project implements full subset regression, Lasso regression, and stepwise regression to train a multi-linear regression to predict the mean of methane. Besides, by deleting some influential points and outliers, the model gets higher accuracy. Finally, the model passes some hypothesis tests, which means the model satisfied the assumptions of muti-linear regression. Finally, I get a well-trained model that can give a satisfactory prediction result with 97% accuracy.

OBJECTIVES

The objective of this study is to predict methane emission and quantitatively evaluate factors that affect methane emission, which could add data-driven insights into climate change mitigation and environmental regulation.

Since the multiple linear regression model simulates the relationship between the dependent variable and multiple independent variables and in the presence of time series, the multiple linear regression model will perform well. Therefore, I use methane emissions as the dependent variable in this research, which not only allows us to determine what factors affect methane levels in the air through the model results but also allows us to predict the average of methane levels in the air over the next period of time.

In addition, the measurement data from the observation sites can have some bias due to chance factors with a relatively small probability. This makes the data inaccurate and thus affects the accuracy of the trained model. Therefore, after the final model is obtained, outliers tests and high-leverage tests are needed to determine the outliers. These outliers are then removed to better improve the prediction accuracy of the model.

METHODOLOGY

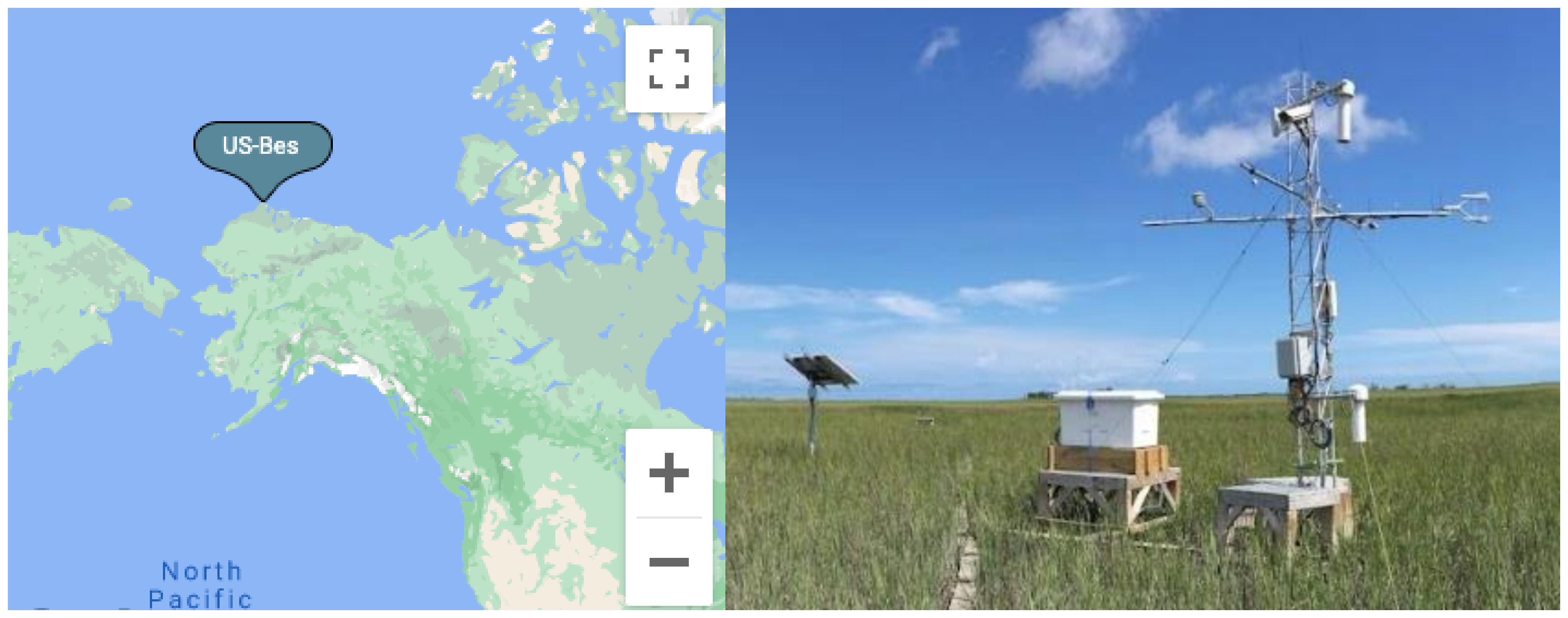

Figure 1. The study site is located about 10 km east of the town of Barrow, Alaska.

The study site shown in Figure1. is located about 10 km east of the town of Barrow, Alaska. The data in this research are all from this study site. This research trained the multiple linear regression model with 1100 data obtained from the study site in the past three years. After getting the final model by stepwise regression etc., I delete some influential points and outliers to improve the model's accuracy. Besides, I draw the Quantile-Quantile Plots and find that the final model satisfied the normality of the errors test, which means probability distribution generally follows a normal distribution. What's more, the probability distribution generally follows a normal distribution. I do the independence of errors test and the homoscedasticity test. Finally, I find that the model satisfied the hypothesis test of linear regression.

RESULTS & DISCUSSION

Firstly, I use "FCH4_mean" as the independent variable and use the remaining data as independent variables to train the full model. Then I get a model with R-squared value about 95 and AIC value about -1207.88. Then I want to increase the R-squared value and decrease the AIC value. Therefore, secondly, I use full subset regression and stepwise regression to select significant variables. As thus, the AIC value of the current model decrease to -1215.58. From the result of the subset regression image (Figure2), I choose the variables to display with black at the top, which means the model formed by these points has the lowest Cp-value and highest adjusted R square.

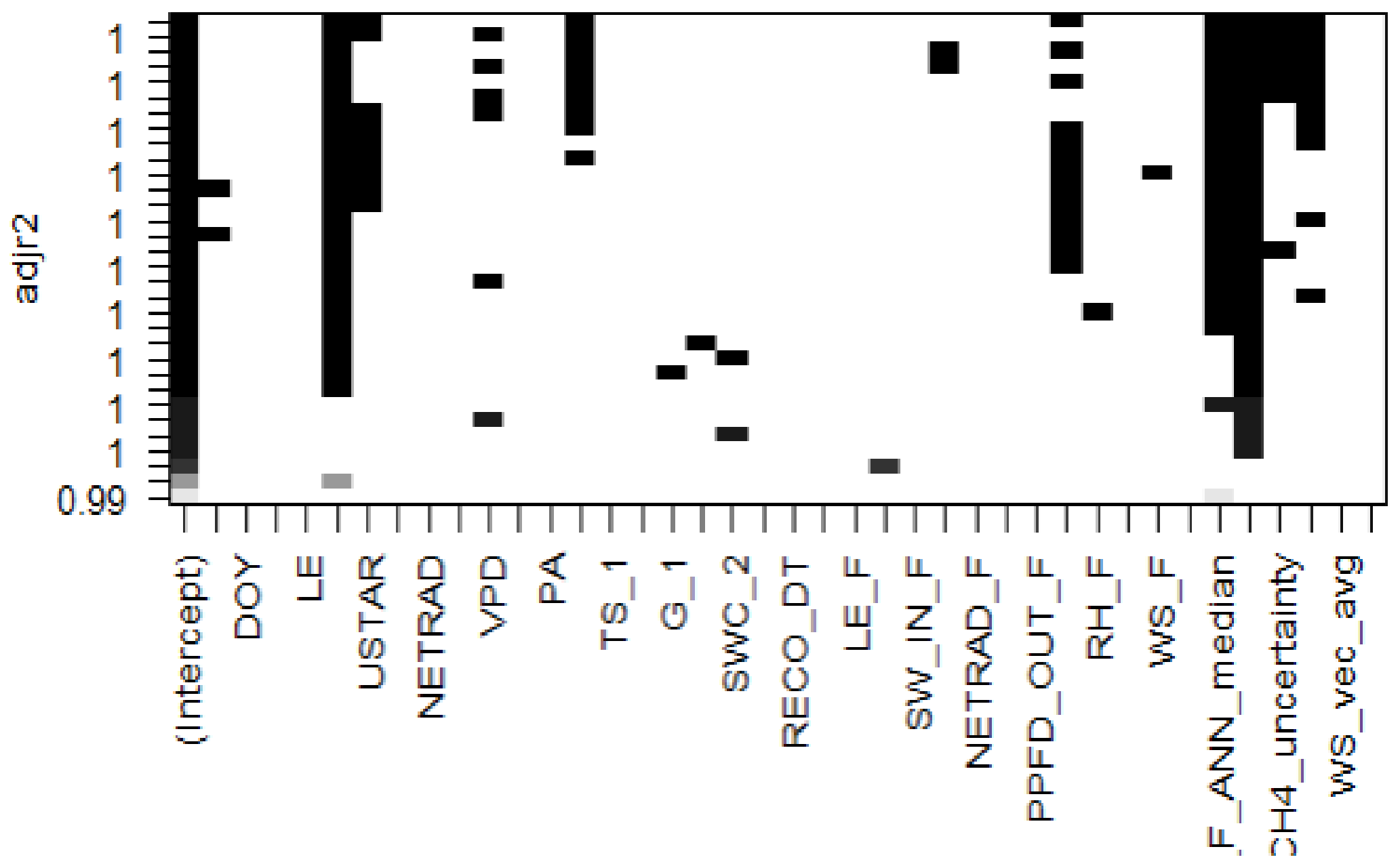

Figure 2. The result of full subset regression

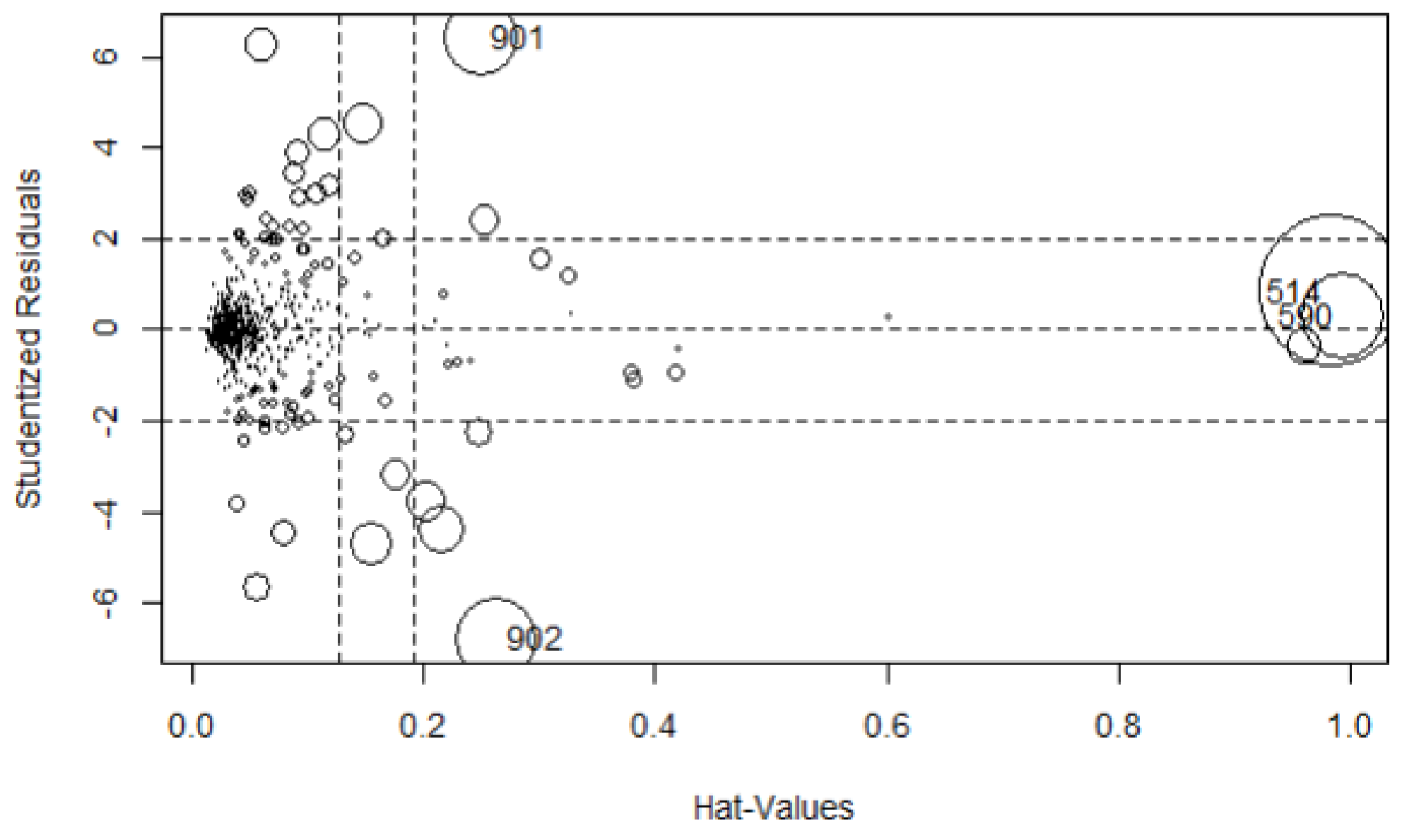

Figure 3. The result of Outliers' plot

Then according to the outliers' test, I calculate the R-student value of each piece of data. And I visualized the results of the R-student value table. The points' index is shown in Figure3,which represents the outliers and influential points, which should be deleted to improve the model performance. Until now I get a well-performed model, then I check the assumptions of linear regression model.

As shown in Figure 4, Cook's distance is the measure of how much the residuals of all cases would change if a particular case were excluded from the calculation of the regression coefficients. Therefore, I finally delete the points, which are considered biased observations, shown in Cook's plot.

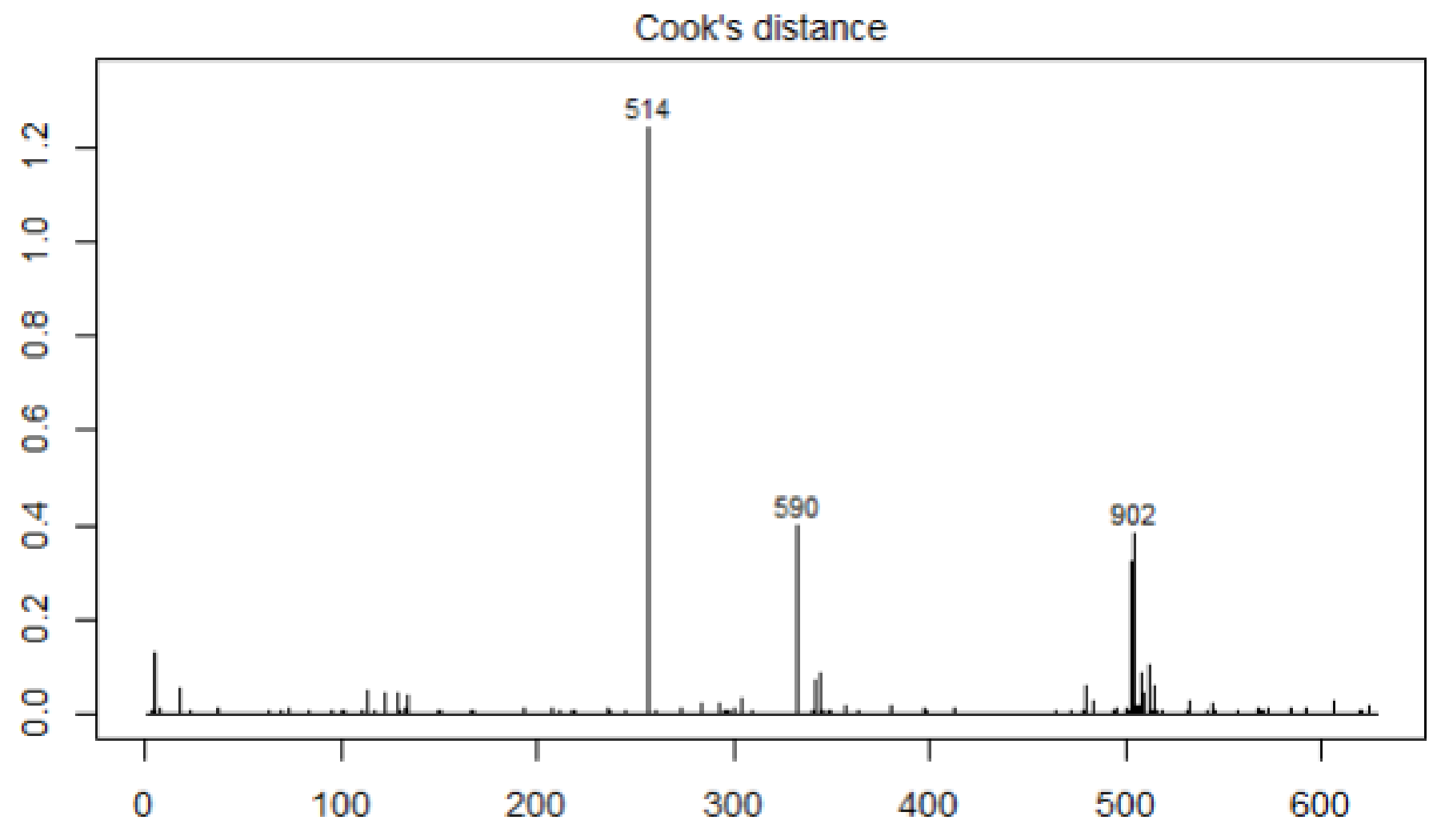

Figure 4. The plot of Cook's distance

RESULTS & DISCUSSION

The following are some hypothesis tests. Firstly, for the independence of errors test, the null hypothesis is that no correlation among the residuals. The DW=2.0045 and p-value=0.5735 > 0.05, errors can be considered independent of each other. The final model satisfies the independence of errors test. Secondly, for the homoscedasticity test, the null hypothesis is that the variance is constant. The p-value of the final model is 0.053>0.05. The model's variance is constant. The final model passed the homoscedasticity test.

Thirdly, the model passes the normal distribution test by drawing the Quantile-Quantile Plots. We know that to identify whether the sample data is approximately normally distributed using the QQ plot, we simply need to see if the points on the QQ plot are approximately near a straight line. The graph is a straight line indicating a normal distribution, and the slope of that line is the standard deviation, and the intercept is the mean.

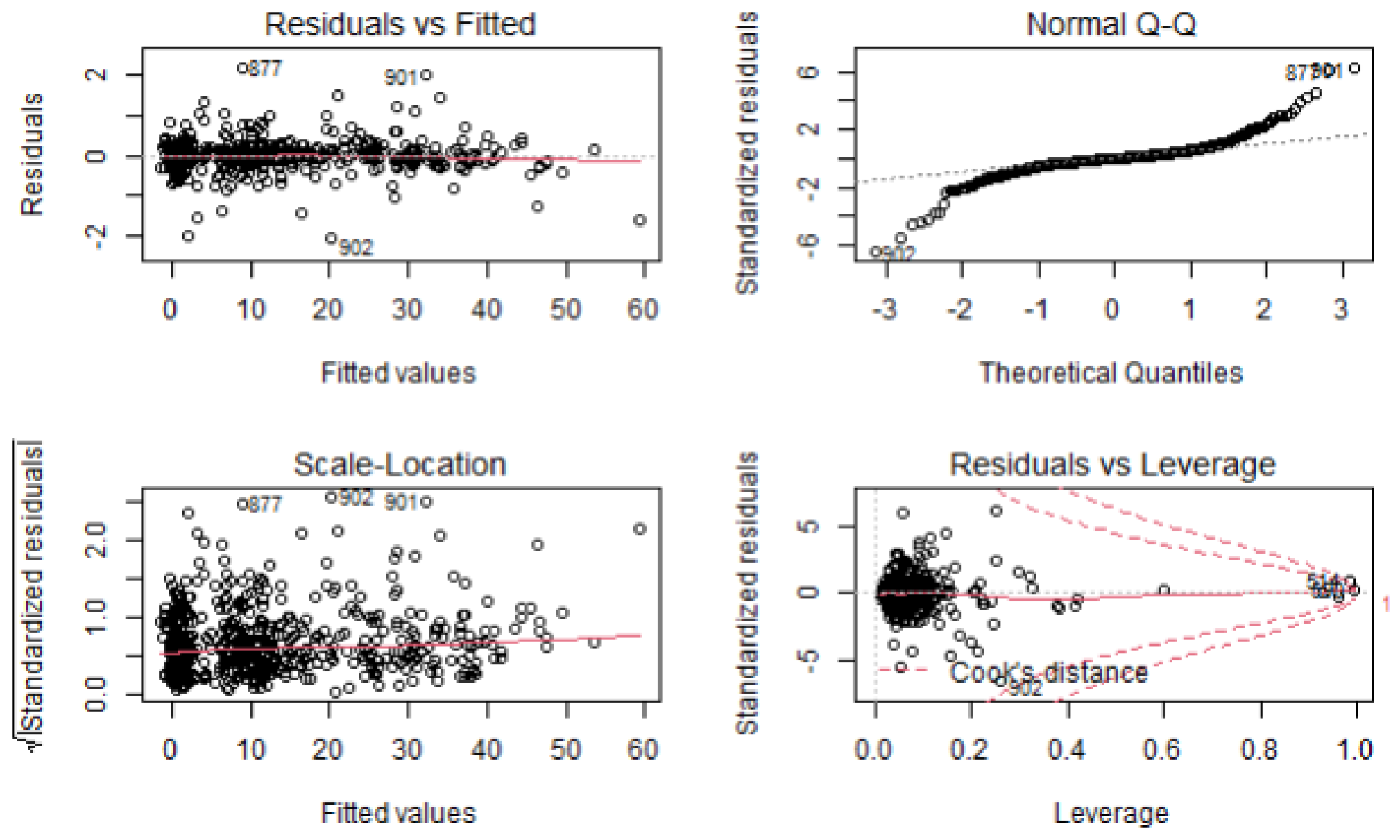

Figure 5. The Q-Q plots of final model

As Figure5. shown above, the plot in the upper right corner shows that the sample data basically obey a normal distribution. For the top left graph, we conclude that the distribution of the residuals essentially follows a normal distribution because the red line in the graph is relatively flat.

I trained a well-performed multilinear regression model with high adjusted R square and low AIC values, which means we can predict the amount of methane in the atmosphere. we can take measures to control the amount of the variables in the model that strongly affect methane emissions in time. This can effectively and directly reduce the amount of greenhouse gases.

ACKNOWLEDGMENTS

This research was supported Dr. Qing Zhu, who has carefully guided me and put forward many suggestions.