Regression-based Analysis of Ozone Layer via Machine Learning Models

Yiran Dong *

Department of mathematics, Purdue University, Indiana, United States

* Corresponding author email: 101048@yzpc.edu.cn

Abstract. Ozone protects the livings on the earth from ultraviolet radiation. The emission of ozone depleting substance causes the Antarctic ozone hole and reduces the ultraviolet radiation absorption rate by ozone layer. Researchers find that ozone depleting substances (ODS) accounts for ozone concentration between 9 km and 25km, by chemical-climate models and ozone concentration will increase by 15% until 2050. In this work, we used six major ODS consumption worldwide and mean stratospheric ozone concentration each year. Seven regression models are implemented to make prediction and k-fold cross validation is used for avoiding overfitting. Root mean squared error (RMSE), and standard deviation are two performance metrics of regression models. The results indicate that the prediction from support vector regression achieved the lowest RMSE. Random forester and k-nearest neighbor are also appropriate for make prediction. We also concluded that linear, polynomial, ridge, and lasso regression methods are hardly to fit the data in this application.

Keywords: Ozone Hole; Machine Learning.

1. Introduction

The natural environment and livings on the earth relies on the Ozone heavily [1]. It protects any living organisms at the Earth's surface against the harmful solar ultraviolet radiation b (UVB) and ultraviolet radiation c (UVC) and serves as an energy budget by absorbing the solar UV and terrestrial IR radiation. UVB is responsible for the occurrence of human skin cancer, sickness of human eyes, and suppression of the human immune system. In addition, ozone in the troposphere has a great impact on warming, evaporation rate, and cloud formation. In 1974, Stolarski and Cicerone [2] proposed that chlorine compounds would deplete the ozone layer by chain reaction. In the reaction, chlorine compounds act as a catalyst, which boosts the formation of oxygen gases by consuming ozone and oxygen atoms. Since the 1950s, the emission of ozone-depleting substances keeps increasing, leading to the thinning ozone layer over the Antarctic. In 1979, scientists observed the Antarctic ozone hole and it keeps expanding in the following years. Thus, the Vienna Convention for the Protection of the Ozone Layer was signed in 1985. The countries involved started to execute the convention in 1988. From 1989, ozone-depleting substance emissions decreased rapidly, which is illustrated in Fig. 1. Until 1991, the area of the Antarctic ozone hole stopped increasing and started to dwindle from 2007. The ozone in the lower stratosphere is affected by the atmospheric circulation and chemical interaction involving anthropogenic chlorine and bromine [3, 4].

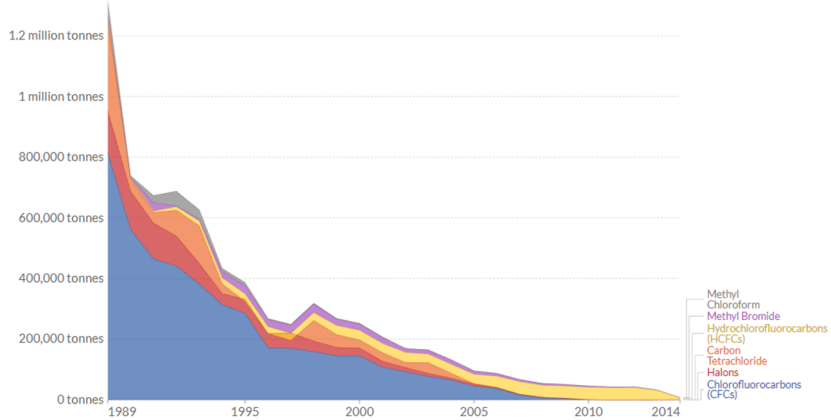

Fig 1. Ozone depleting substance (ODS) consumption, worldwide [5].

Several previous studies also focused on ozone problems, primarily using time series and chemistry-climate models, like UMCHEM, equivalent effective stratospheric chlorine (EESC), and the University of L' Aquila model (ULAQ model). Details are included in Section 2.

In this paper, we analyze the stratospheric ozone concentration and ODS consumption worldwide from Our World in Data [5] to investigate the performance of seven regression models, including linear regression, lasso regression, ridge regression, random forester regression (RFR), polynomial regression, k-nearest neighbors' regression (KNN) and support vector regression (SVR).

2. Related Works

We reviewed several works about this direction. Eun-Su Yang in 2006 used time series, equivalent effective stratospheric chlorine (EESC), and sophisticated photochemical calculations to analyze the data obtained from TOMS/SBUV satellite, 36 Dobson/Brewer stations, and SAGE and HALOE satellite [6]. They concluded that both ozone and stratospheric halogen abundances reached their peak around midway through 1997. The increases in stratospheric ozone at northern hemisphere midlatitudes since 1997 are consistent with the ozone increase for altitudes below 18 km, probably resulting from changes in atmospheric dynamics.

Heterogeneous chemical reactions involving sulfate and polar stratospheric cloud (PSC) aerosol surfaces are also factors leading to the ozone losses. Pitari investigated the relative humidity, cloud distribution, and net precipitation rates taken from climatological data, that is, Oort [7] and Shea [8] by the University of L'Aquila model (ULAQ model) in 2002 [9]. Both climate and ozone depleting substance emission changes will restrict the ozone recovery. 2/3 of the delay of the O^3 recovery rate results from the climate changes, including perturbed stratospheric circulation, nitric acid distribution, and PSC frequency.

3. Methodology

3.1 Data Description

Table 1. General description of ODS and ozone range, distribution.

ODS	Methyl Bromide	Carbon Tetrachloride	HCFCs	CFCs	Methyl Chloroform	Halons	Ozone
count	18	18	18	18	18	18	18
mean	13131.0	46935.0	23054.0	219757.8	10288.3	42242.2	111.3
std	9715.0	71312.9	11986.9	228768.2	16906.8	46459.1	10.9
min	127.9	0.2	2033.3	155.5	0	0.1	92.3
25%	5516.6	1237.5	13223.2	50393.8	663.6	5812.9	101.4
50%	12025.2	30488.8	26783.4	144759.5	1925.5	26659.9	113.2
75%	21382.8	43990.9	30681.8	365102.2	8438.8	62077.2	118.3
max	29648.5	124463.7	41017.7	815613.0	52069.2	136865.5	128.6

count: the number of training samples of ODS and ozone std: standard deviation

The data used in this research were obtained from Our World in Data [5], a scientific online publication. Our World in Data, based on the University of Oxford, concentrates on the environment, health, poverty, education, and demographic change. The feature data are the amount of six ozone-depleting substances (ODS) from 1989 to 2014, including methyl bromide, carbon tetrachloride, hydrochlorofluorocarbons (HCFCs), chlorofluorocarbons (CFCs), methyl chloroform, and halons, multiplied by their respective ozone depleting potential value. These six ODS are the major pollutant released by human acclivities. Halon and CFCs are the most harmful to the stratosphere ozone layer. Therefore, the feature data would be a good representative of factors, resulting from human activities. About 90% of the planet's ozone is in the stratospheric layer, and most ozone decomposition reactions

happened in the upper stratospheric layer. The mean daily stratospheric ozone concentration, instead of the Antarctic ozone hole area, is used to describe the ozone layer. Detailed information about the data can be found in Table 1.

3.2 Data Preprocess

The concentration of six ODS multiplied by their ozone depleting potential value is set as feature variables, while the stratospheric ozone concentration is the target value. The target value was once set as the difference in stratospheric ozone concentration, but the RMSE and standard deviation were larger. The ODS production for several years is negative. The data for these years are removed from the training samples.

3.3 Algorithms

3.3.1 Support Vector Regression

Suppose a set of training data $\{(x_1, y_1), (x_2,y_2)...(x_m, y_m)\}$ is given, where $x_i \subseteq X \subseteq R^n, y_i \subseteq R$ and m is the number of training samples. Support vector regression optimizes the function f(x) such that all the training data is within ε deviation, while support vector regression does not overfit the training data [10]. The support vector regression is presented as

$$f(x) = < w, \phi(x) > + b \tag{1}$$

where w is the weighting coefficient vector related to ϕ(x), ϕ(x) maps x to a vector in F a higher-dimension feature space, b is the bias term and <, > represents inner product. To deal with the situation that the training data is not linearly separable, the kernel trick is applied here. Kernel functions are similar functions that salsify Mercer's theorem such that

$$k(x_i, x_i) = < \phi(x_i), \phi(x_j) > \tag{2}$$

Moreover, there does not always exist a function f (x) that approximates all (xi,yi) within ε precision. Slack variable ξ and ξ* is used to measure the errors larger than ε. While the penalty function is linear, support vector regression is optimized by the following equations.

$$\text{minimize } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i + \xi_i^*$$

$$\text{subject to} \begin{cases} y_i - < w, \phi(x_i) > - b \le \varepsilon + \xi_i \\ < w, \phi(x_i) > + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{3}$$

where the first term represents the model flatness, preventing the function from overfitting.

The second term measures the sum of positive and negative training errors larger than ε. The constant C > 0 determines the tradeoff between model flatness and training errors. Lagrange function is applied to solve the objective equation (3) and a saddle point is proved to exist by Mangasarian [11], McCormick [12], and Vanderbei [13].

3.3.2 K-nearest Neighbor Regression (KNN)

K-nearest neighbor regression (KNN) forecasts the number of predicted variables by the training data with feature similarity. In KNN, there is no equation between the predicted variable and predictor. Feature similarity is determined by the distance between the predicted variable and training variables. The common calculation methods are Manhattan (p=1), Euclidian (p=2), and Minkowski distance function (p is arbitrary) as below:

$$D_{rt} = (\sum_{i=1}^{m} |x_{ir} - x_{it}|^p)^{\frac{1}{p}} \tag{4}$$

The predictor equals the summation of k-nearest training data multiplied by the weights of each point ($w(D_{rt})$):

$$y_r = \sum_{j=1}^{K} w(D_{rj}) \times y_j \tag{5}$$

In our experiment, weight functions do not show significant difference, so we are using uniform weight function, which is:

$$w(D_{rj}) = \frac{1}{K} \tag{6}$$

3.3.3 Random Forester Regression

Random forester regression uses the unweighted average of a large number of decision trees, also called Classification and Regression Tree (CART) to make prediction. A decision tree is composed of decision nodes and leaf nodes. The input samples are evaluated at decision nodes by a split function. It is then and sent to different branches depending on feature of the samples and arrives at leaf node in the end. Suppose the observed data is S= $\{(x_1, y_1), (x_2, y_2)...(x_n,y_n)\}$, where $x_i \subseteq R^p, y_i \subseteq R$ and p is the number of features. Random forester regression is a collection of tree predictors $h(x, \theta_i)$, i=1, 2, ..., T, where θ_i are independently and identically distributed. It randomly selects n samples with replacement from D. All samples have equal probability of 1/n to be chosen. The process above, called 'bootstrapping', is repeated T times and T sub-datasets are generated ($\theta_1, \theta_2, \ldots, \theta_T$). A collection of q independent learner ($h_1(x, \theta_1), \ldots, h_T(x, \theta_T)$) are trained based on each sub-dataset. The prediction is made by averaging the outputs of each base learner.

$$\hat{h}(x) = \frac{1}{T}\sum_{i=1}^{T} h_i(x, \theta_i) \tag{7}$$

The algorithm above, called 'bagging' cannot only be applied to training data, but also used for features. It is an effective way to minimize the variance of the prediction, by preventing the prediction from overfitting to each base leaner.

3.3.4 Linear Regression

Linear regression is a linear function of feature variable:

$$\hat{y_i} = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} \tag{8}$$

$X_i = \{x_{ij}, j = 1,2, \ldots m\}$ represents the feature set from i^{th} data. $\hat{y_i}$ is the predicted target value of the i^{th} data. β_0 is the intercept term and β_i is known as regression coefficient. In linear regression β is obtained by

$$\text{minimize: } \sum_{i=1}^{n}((\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}) - y_i)^2 \tag{9}$$

where y_i is the observed target value of the i^{th} data.

3.3.5 Lasso Regression

Lasso regression is a type of linear regression that performs L1 regularization. L1 regularization penalize the regression model with the absolute value of coefficients, leading to simple, sparse models. Some coefficients may be eliminated because of the penalty. Lasso regression has the same function as linear regression (9). The only difference is calculation of β:

$$\text{minimize: } \sum_{i=1}^{n}((\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}) - y_i)^2 + \alpha \sum_{j=1}^{m} |\beta_j| \tag{10}$$

α, a tuning parameter, regulates the strength of the L1 penalty. As α increase, more coefficients are eliminated.

3.3.6 Ridge Regression

Ridge regression performs L2 regularization, which adds a penalty term, equivalent to the square of coefficients. It only shrinks coefficients by the same factor, instead of eliminating coefficients. Ridge regression has the same function as linear regression and lasso regression (10). The difference is also the calculation of β:

$$\text{minimize: } \sum_{i=1}^{n}((\beta_0 + \sum_{j=1}^{m}\beta_j x_{ij}) - y_i)^2 + \alpha\sum_{j=1}^{m}\beta_j{}^2 \tag{11}$$

3.4 Methods Optimization

Since the data of ODS production started to be collected from 1989, the sample sets are limited. We are using K-fold cross validation to evaluate the effectiveness of the models, while setting k to five.

3.5 Performance Metrics

3.5.1 Standard Deviation (SD)

Standard deviation describes the variation of a set of values. In this experiment, we used k-fold cross validation method. Standard deviation of root mean squared error reflects the variation of model performance of each fold

3.5.2 Root Mean Squared Error (RMSE)

RMSE is a measure of the predicting variation. It has similar formula as standard deviation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{12}$$

where n is the number of observations, y_i is the observed values, and $\hat{y}_i$ is the predicted values. RMSE describes the accuracy of the prediction. In general, the smaller the RMSE is, the more accurate the prediction is. Moreover, RMSE is sensitive to outliers.

4. Results

4.1 Experimental Setting

To compare the efficiency of models, the result of each model should be optimized. Parameters, like the regularization term, may be influential. We adjust these parameters of each model and grade the parameters based on the RMSE and standard deviation of the test.

For lasso regression, we alter the constant that multiplies the L1 term (α). We set α to 0.1, 1, 10, 100, and 1000. The corresponding RMSE are 25.04, 25.03, 24.87, 23.37, 14.92 and the corresponding standard deviation are 32.84, 32.81, 32.51, 29.48, 12.1. In Fig. 2, RMSE and standard deviation of α values less than 100 are about the same, while RMSE and standard deviation of α equivalent to 1000 decrease obviously. Consequently, we determine 1000 to the regularization constant of lasso regression.

α in ridge regression is the constant that multiplies the L2 term. However, when α of ridge regression is set to 0.1, 1, 10, 100, and 1000, the RMSE and standard deviation are around 25 and 33. The regularization constant of ridge regression has little effect on RMSE and standard deviation, so in the ridge regression model, α is 1000.

As for support vector regression, we tune the regularization term (C), and ε in the ε-SVR model. The kernel is another common parameter of support vector regression, but the radial basis function (RBF) kernel undoubtedly outperforms the linear kernel and polynomial kernel. Therefore, the RBF

kernel is the default below. As shown in Table 2, RMSE is mostly determined by regularization terms and reaches a minimum when C equals 300. The standard deviation of support vector regression decreases, as C increases. Considering the standard deviation is small in all cases, C is set to 300 and ε to 0.1.

Fig 2. Performance metrics of lasso regression with α=0.1, 1, 10, 100, 1000

The number of neighbors (n), and power parameters for the Minkowski metric (p) are the two parameters adjusted in the k-nearest neighbors regression model, but they hardly affect the performance of the model. RMSE is around 12 and SD is 4, so the setting (p=2, n=5), with the minimum RMSE, is selected.

Table 2. The parameters optimization of SVR based on RMSE and standard deviation
C: regularization term, ε: ε in equation 3

RMSE	Ɛ=0.1	Ɛ=0.2	Ɛ=0.3	Ɛ=0.4
C=0.1	12.29	12.29	12.31	12.33
C=10	10.19	10.24	10.31	10.37
C=100	7.59	7.68	7.78	7.86
C=300	7.18	7.20	7.23	7.28
C=400	7.28	7.32	7.35	7.37

SD	Ɛ=0.1	Ɛ=0.2	Ɛ=0.3	Ɛ=0.4
C=0.1	1.13	1.15	1.16	1.17
C=10	1.60	1.61	1.63	1.65
C=100	2.73	2.65	2.58	2.51
C=300	3.35	3.36	3.36	3.35
C=400	3.54	3.55	3.54	3.53

Table 3. The parameters optimization of KNN based on RMSE and standard deviation

RMSE	p=1	p=2	p=3
n=3	11.94	12.16	12.37
n=4	11.63	12.22	12.22
n=5	11.81	11.47	11.47
n=6	12.15	12.26	12.26

SD	p=1	p=2	p=3
n=3	4.39	3.88	3.18
n=4	3.97	3.55	3.55
n=5	4.07	4.00	4.00
n=6	4.12	4.01	4.01

Table 4. The parameters optimization of RFR based on RMSE and standard deviation

RMSE	f=1	f=2	f=3	f=4
n=100	10.57	10.76	11.25	11.37
n=500	10.9	11.18	11.15	11.41
n=1000	11.03	11.35	11.34	11.26

RMSE	f=1	f=2	f=3	f=4
n=100	2.76	3.58	4.07	4.56
n=500	2.72	3.12	3.65	3.77
n=1000	2.77	3.00	3.21	3.61

RMSE and standard deviation of polynomial regression both enhance, as the degree of the polynomial (n) increases. Polynomial in this case is simplified to linear regression.

We tune the number of trees in the forest (n), and the number of features during the split (f). Table 4 demonstrates that both RMSE and standard deviation have the best performance when the number of trees is 100 and f is one.

4.2 Performance

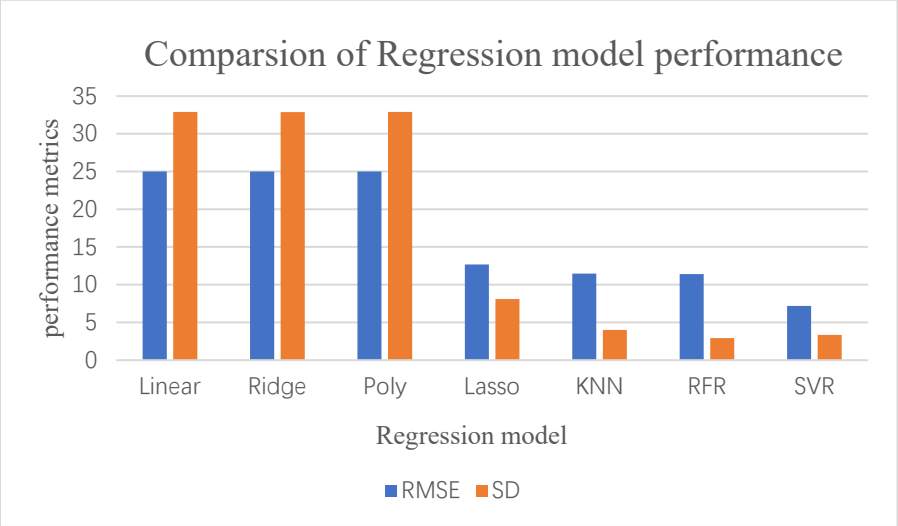

Fig 3. The result of testing models application to six major ODS consumption worldwide

SVR outperforms other models in this experiment. According to Fig. 3, both RMSE and standard deviation of SVR are the lowest among all the models. SVR model not only has high accuracy but also shows low dispersion of the prediction. The mean stratospheric ozone concentration of training data is 111.3, while the RMSE of lasso regression, KNN, and RFR are a little bit higher than 10, around 10% of the mean stratospheric ozone concentration, so the RMSE of these three models are acceptable. However, the standard deviation of lasso regression is obviously larger than the other two models. KNN and RFR would make a more stable prediction than lasso regression. The performance of linear, ridge and polynomial regression is the same. Their RMSE value is about 25 and the standard deviation is over 30. Their prediction is invalid and has a high variance. Therefore, they cannot be applied to predicting the stratospheric ozone concentration.

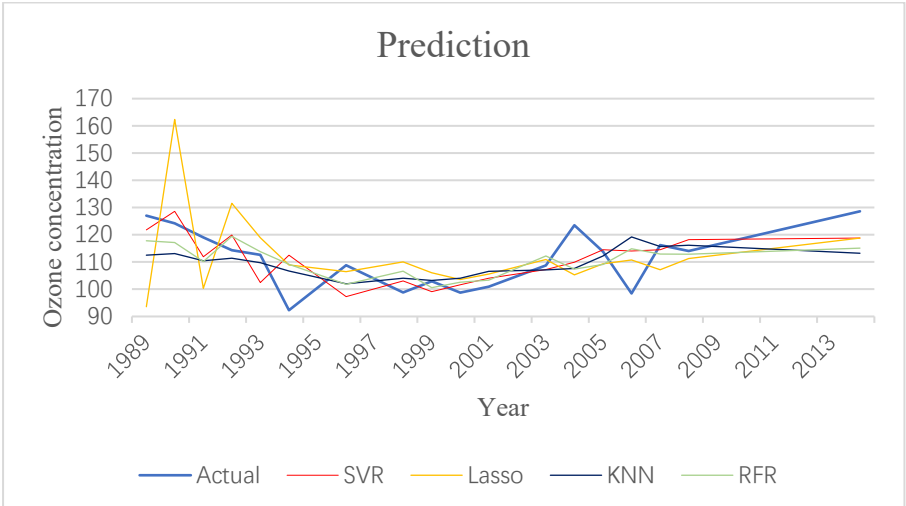

Fig 4. The prediction of appropriate models and the actual stratospheric ozone value. K-fold cross validation is not applied in the prediction. All the other data is used for training the models

SVR, KNN, and RFR make similar predictions. Fig. 4 shows that their predicted ozone concentration declines from 120DU to 100DU and starts climbing back to 120DU. The turning points of these three models are all around 2000. The trend predicted by lasso regression is almost the same as the trend by SVR, but the fluctuation of lasso regression is extreme, especially from 1980 to 1993. Observation confirmed the predictions in most cases. but in 1994, 2004, and 2006, the actual ozone concentration is absolutely different from all the forecast values.

4.3 Discussion

In this paper, linear regression methods, like linear, ridge, lasso, and SVR with linear kernel, do not perform well in predicting the stratospheric ozone concentration, probably resulting from the nonlinear relationship between the production of ODS and ODS reacting to ozone in the stratosphere layer. We use six major ODS as feature variables, but apart from anthropogenic ODS, natural ODS, solar radiation, and atmospheric transport also affect stratospheric ozone. These factors could be further used for modeling.

This experiment is restricted by the number of data sets. The regression models tend to overfit the model, which is demonstrated in the part of parameter setting. The regression models achieve better results with high regularization terms. The anthropogenic ODS was introduced in 1970s and the consumption of ODS has been recorded since 1989. It is not likely to increase the size of the data. The experiment could be developed by eliminating less important ODS or using a dimension reduction method, like the principal component analysis.

5. Conclusion

In this experiment, we successfully evaluate the performance of seven regression models. SVR performs the best in forecasting the ozone layer among seven regression models. It has the highest accuracy and lowest variance of the prediction. KNN and RFR are appropriate models for the prediction, but have larger RMSE than the SVR model. Linear, ridge, lasso, and polynomial regression are not suitable for the prediction.

References

[1] Cicerone, R. J. (1987). Changes in stratospheric ozone. Science, 237(4810), 35-42.

[2] Stolarski, R. S., & Cicerone, R. J. (1974). Stratospheric chlorine: a possible sink.

[3] World Meteorological Organization (1999), Scientific assessment of ozone depletion: 1998, 44 pp., Geneva, Switzerland.

[4] World Meteorological Organization (2003), Scientific assessment of ozone depletion: 2002, Global Ozone Res. Monit. Proj., Rep. 47, Geneva, Switzerland.

[5] Hannah Ritchie and Max Roser (2018) - "Ozone Layer". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/ozone-layer'.

[6] Yang, E. S., Cunnold, D. M., Salawitch, R. J., McCormick, M. P., Russell III, J., Zawodny, J. M., ... & Newchurch, M. J. (2006). Attribution of recovery in lower-stratospheric ozone. Journal of Geophysical Research: Atmospheres, 111(D17).

[7] Oort, A. H. (1983). Global atmospheric circulation statistics, 1958-1973 (No. 14). US Department of Commerce, National Oceanic and Atmospheric Administration.

[8] Shea, D. J. (1986). Climatological Atlas. National Center for Atmospheric Research.

[9] Pitari, G., Mancini, E., Rizi, V., & Shindell, D. T. (2002). Impact of future climate and emission changes on stratospheric aerosols and ozone. Journal of the Atmospheric Sciences, 59(3), 414-440.

[10] Vapnik, V. N. (1995). The nature of statistical learning. Theory.

[11] Mangasarian, O. L. (1969). Nonlinear Programming “McGraw-Hill Book Company”. New York.

[12] McCormick, G. P. (1983). Nonlinear programming: Theory, algorithms, and applications. John Wiley & Sons, Inc.

[13] Vanderbei, R. J. (1999). LOQO user's manual--version 3.10. Optimization methods and software, 11(1-4), 485-514.