Indonesian Journal of Electrical Engineering and Computer Science
Vol. 17, No. 3, March 2020, pp. 1461~1466
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v17.i3.pp1461-1466

Ozone prediction based on support vector machine

M. Tanaskuli[1], Ali Najah Ahmed[2], Nuratiah Zaini[3], Samsuri Abdullah[4], Abdoulhdi A. Borhana[5], N. A. Mardhiah[6], Mathivanan[7]

[1,2,3,7]Department of Civil Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), Malaysia
[2]Institute of Energy Infrastructure-IEI, Universiti Tenaga Nasional (UNITEN), Malaysia
[4]Air Quality and Environment Research Group, Faculty of Ocean Engineering Technology and Informatics, Universiti Malaysia Terengganu
[5,6]Department of Mechanical Engineering, College of Engineering Science & Technology, Sebha University, Libya
[5]Department of Mechanical Engineering, College of Engineering, Universiti Tenaga Nasional (UNITEN), Malaysia

Article Info

Article history:

Received Feb 23, 2019
Revised Apr 12, 2019
Accepted Jul 5, 2019

Keywords:

Klang valley Malaysia
Ozone concentration
SVM

ABSTRACT

The prediction of tropospheric ozone concentrations is very important due to negative effects of ozone on human health, atmosphere and vegetation. Ozone Prediction is an intricate procedure and most of the conventional models cannot provide accurate prediction. Machine Learning techniques have been widely used as an effective tool for prediction. This study is investigating the implementation of Support vector Machine-SVM to predict Ozone concentrations. The results show that the SVM is capable in predicting ozone concentrations with acceptable level of accuracy. Sensitivity analysis has been conducted to show what is the most effective parameters on the proposed model.

Corresponding Author:

Ali Najah Ahmed,
Institute of Energy Infrastructure-IEI,
Department of Civil Engineering,
Universiti Tenaga Nasional,
Jalan Ikram-Uniten, 43000 Kajang, Selangor, Malaysia.
Email: Mahfoodh@uniten.edu.my

1. INTRODUCTION

Dissolved Air pollution has been the main problem to be concerned in a developed and developing countries [1-3]. Ozone layer is an optional photochemical toxin created from an assortment of common and anthropogenic antecedents that incorporate mechanical and vehicular emanation of unstable natural mixes and oxides of nitrogen [4]. Ozone is an imperative constituent of the air that keeps up the warm structure of the stratosphere and the troposphere [5]. Numerous studies focused on prediction ozone concentrations [6]. These proposed models are not accurate enough. Many researchers focusing now on using Artificial Intelligence techniques such as support vector machine in replacing the conventional method in prediction [7-12]. Therefore, the main objective of this study is to investigate the capability of SVM in predicting Ozone in three different locations in Malaysia. Hyper parameters optimization will be investigated to improve the accuracy of the proposed model. Then, finally sensitivity analaysis will be introduce to test the importance of each input on the proposed model.

2. METHODOLOGY

2.1. Study Area and Data

Nine air quality monitoring stations were selected for this study as shown in Table 1 namely; Klang (S1), Petaling Jaya (S2), and Cheras (S3). Cheras is suburb of Kuala Lumpur, the capital city of Malaysia. The township is located to the south-east of Kuala Lumpur. The air monitoring station at Sekolah Menengah Kebangsaan Sri Permaisuri with coordinate at (3°06'25.2"N 101°43'03.9"E).Petaling Jaya, Selangor is geographically located at coordinate (3°08'16.7"N 101°36'29.7"E) is located within Klang Valley region and covers an area of 97.2 km^2. Its location near to Kuala Lumpur's city centre, surrounded by residential, commercial and industrial areas leads to a high volume of traffic. The monitoring station for Petaling Jaya was located at Sekolah Rendah Bandar Utama, Petaling Jaya with coordinate (3°08'16.7"N 101°36'29.7"E).

In order to conduct the prediction ozone layer concentration a three set of data were collected from the Department of Environmental Malaysia. Table 2 represent the minimum, maximum, average and correlation of the date set of Cheras. In this data, it contains 7 parameters which are Wind Speed hourly, humidity hourly, NOx Hourly average, So2 hourly average, No2 hourly average, O_3 hourly average and Co hourly average.

Table 1. Air Quality Monitoring Stations Description

Air Monitoring Station	Location	Background	Coordinates	
			Latitude	Longitude
S1	Klang	Industrial	3°6.'20.0"N	101°24'48.4"E
S2	Petaling Jaya, Selangor	Industrial	3°08'16.7"N	101°36'29.7"E
S3	Cheras, KL	Urban	3°06'25.2"N	101°43'03.9"E

Table 2. Analysis of Parameters

Parameters	Min	Max	Avg	R
Ws hourly avg	0.8	1.5	1.15	0.716394
Humidity hourly avg	100	100	100	-0.74198
NOx hourly avg	0.031	0.033	0.032	-0.51264
So2 hourly avg	0.002	0.004	0.003	0.17301
No2 hourly avg	0.024	0.028	0.026	-0.32457
O3 hourly avg	0.002	0.014	0.008	1
Co hourly avg	0.57	0.81	0.69	-0.47499

2.2. Regression in Support Machines

An SVM comprises network architectures similar to that of ANNs, which have been pruned in order to acquire model simplicity or enhance generalisation. Nevertheless, the approaches for determining network architectures regarding either model are dissimilar. Determining suitable ANN architectures typically requires manual procedures involving trial and error that mostly depends upon the past experiences and preferences of users, with derived weights that cannot be interpreted [13-15]. On the other hand, SVM network architecture is analytically confirmed through SVM algorithms, with optimised support vector networks obtained as a result.

3. RESULTS AND DISCUSSION

3.1. Parameter and Characteristic of SVM

To discover the proposed model for the study, we had to go through SVM to find out the best kernel function for the study. Examples of kernel functions are Radial Basis Function, linear function, polynomial, and sigmoid function. The main reason of the findings is to come out with the best kernel functions that could produce an excellent result. Besides that, digit parceling preparing information is aligned by applying the SVMs with various portions to make the last models engineering. Table 3 presented a similar examination for expectation outcome of the S total SVM display utilizing four kernel functions. It could be watched that concerning connection coefficient values (R2) while utilizing the RBF bit, the expectation exactness of the test session information ended up being the best with around 0.86, trailed by polynomial (0.853), the linear kernel (0.77), and the sigmoid kernel (0.062).

In real scenario, finding out for a good SVM model probably the first things to be done for the best structure of SVM model for specific application, there are two indispensable parameters that expected to be chosen in particular; limit parameter C and ε [16-18]. The determination of C is exceptionally delicate to the exactness of the forecast, as the little estimation of C could tend the model to under gauge the objective

esteem amid the preparation information, this is because of the way that utilizing generally little weight as a part of the preparation information would mirror a bigger estimation of the indicator while inspecting the model in the testing dataset and vice-versa. On the highest point of that, when C is huge, the weight will lose its centrality on recognizing the mapping between the info and the yield. On the other hand, the huge estimation of C could mirror an extensive variety of support vector's qualities; in like manner, supplementary information records could be decided for enhancing the support vectors. Moreover, substantial estimation of ε could tend for less number of the support vector to be accomplished and afterward the normal representation of the proposed clarification is deficient. Moreover, too substantial estimation of ε could prompt to disintegration of the level of exactness amid the preparation. In this specific situation, the choice of the ideal estimations of both C and ε ought to be accomplished through a few experimentation systems. Once the ideal estimations of these two parameters are recognized, there is a high potential to accomplish abnormal state of precision for foreseeing the fancied information.

Table 3. Outcome from each Kernel Function

Kernel Function	MSE		R	
	Training	Test	Training	Test
Linear	0.000	0.000	0.770	0.768
Polynomial	0.000	0.000	0.839	0.838
RBF	0.000	0.000	0.853	0.854
SIGMOID	0.002	0.002	0.062	0.036

Current review, motivation behind deciding proper estimations for parameter C and ε, the replication forms a few parameters landscapes likewise observed the highlights for SVM S total by way of real stride to conduct expectation display. Toward the starting, a thought of consistent estimation of ε to be 0.1, and variable estimations of C to be gone somewhere around 0 and 10 keeping in mind the end goal to assemble the proposed show amid the instructional course of the info yield information [19-21]. Thusly, a computation of the expectation mistake as RMSE and R2 is completed with recognizing the quantity of the quantity of the support vectors. It could be portrayed in Figure 1 that slight decrease in the quantity of the support vectors and the estimation of estimation of RMSE are accomplished while the utilized estimation of C expanded, then again, the estimation of the R2 increments. Moreover, with concentrating on the parameter C, it could be watched that the most minimal estimation of RMSE point (0.816088231) and one high connection coefficient esteem (0.920869155). Correlation Coefficient pattern increase on point then it drops and increase. Subsequently, it is ideal to choose parameter C to be 2.0.

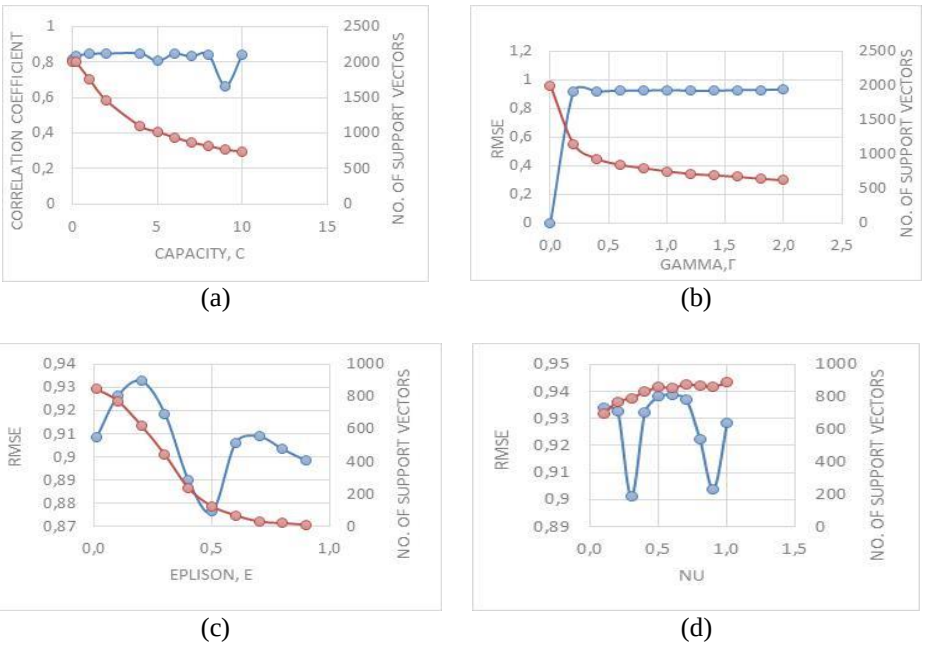

Figure 1. (a) Capacity Vs Correlation coefficient (b) Gamma VS RMSE (c) Epsilon- RBF Model (d) Nu-RBF Model

Really, it is important to make sense of the suitable estimations of the hyper parameters C and γ as a noteworthy stride while actualizing any SVM show [22-24]. In this way, there is a need to complete experimentation methodology. In this unique situation, estimation of the summed-up exactness using different estimation of part hyper has been finished. With deference, the impediment of parameter γ, it was chosen to hunt its ideal esteem to be inside the scopes of [0.20, 2.0] at augmentation of 0.2 for γ with being altered to each C=2 and ε=0.2. Thusly, the ideal determination estimation of γ is discovered utilizing 10-fold cross-approval with dull ten times keeping in mind the end goal to upgrade the unwavering quality of the model results. In the instructional course, the last engineering of the model in the opposition, when such parameters may deliver the base mistakes amid the approval session. Figure 1(b) demonstrates the relationship between RMSE and γ, where it started in the estimation of relationship coefficient ascend with expanding γ until it achieves its pinnacle esteem at 0.4 and after that esteem begins to stay continues alongside in the quantity of support vectors. The estimation of parameters that gives the base speculation blunder is then chosen. The best result for the S total display in preparing and gauging stage while selecting γ=2.0 with a worthy estimation of number of support vectors 622.

For the most part, selecting the ideal number of parameters contributions of a specific SVM model is fundamental stride; notwithstanding, cutting-edge it is hard to discover certain hypothesis that could be utilized to chaperon accomplishing this progression. While playing out the preparation and testing session of SVM model, a similar info course of action of the information set of stone segment parameters that were utilized have been said as a part of the past areas of this paper. Indeed, looking for the model parameters assumes a vital part in accomplishing a decent execution for SVM. Set the hyper parameters C, γ and the portion parameters (Epsilon and Nu) is considered as indispensable stride in impacting on the SVM speculation execution (estimation exactness) [25]. This segment will be centred in the utilization of the two sorts of RBF piece for its great execution and points of interest in ozone concentration determining issue. The accompanying clarifies improved and chose part parameter values for each of the two models.

Epsilon-Radial Basis Function display cast-off to produce a foreseeing S total, here settle C = 2 and γ=2.0, customary ε as different values somewhere around 0.01 to 0.9. Outcomes execution for the perfect in preparing also testing appeared in Figure 1(c), we can watched that the RMSE increment with expanding the estimation of ε, until ε=0.3 while each of the R2 and the quantity of support vectors that are diminished. For the last the ε esteem (0.2) that yields the base speculation mistake with worthy number of support vector (622) is then, picked.

In performing runs with this model, we utilized ideal qualities in every specific gamma= 2.0, limit =6 parameters with an output. The model expands, as forecast precision of the model run information builds bit by bit to the most noteworthy esteem (when Nu=0.6) and afterward the estimation of the rest of the abatement as appeared Figure 1(d) additionally demonstrates the number identified with support vectors increment with the expansion in Nu esteem.

Previous method of 10 - fold validation been used in trial and error method. Also, method of time series regression been tried. As the final testing, V-Fold validation been tested. The cross-approval development remains as one of broadly utilize strategies in assessing the values. With a specific end goal to utilize the approval, preparation of a set of data randomly part in a usual number for V-folds. At that point the chose sort of SVM model is performed successively to the perceptions that occurred to the V-1 folds. Methodology could achieve the eventual outcomes of the acting outline which illustration or wrinkle that was hidden while setting up the SVM appear; i.e., this is the trying example) keeping in mind the end goal to make sense of the mistake characterized by one of the measurable record. The real favourable position of this procedure is that the normal precision for the v times could prompt to steady quantify display mistake on the dependability.For an instance the lawfulness for testing session of subtle data with the model itself. Statistical evaluation total using 8, 12 and 16 as shown in Table 4.

Table 4 outlines that the MAE and RMSE values accomplished using 8-crease cross-approval demonstrate best goodness fitting and extraordinary execution if looked at while using 12 overlay cross-approval. Moreover, the results obtain were for MAPE values. It appears to be basic for the decision to speak to for the component for both the preparation the model and testing session.

Besides that, using Support Vector Machine this research also has been conducted using time series regression. The outcome from the run be identified based on the results. A value from the multilayer prediction model 6-9-1 presents the graph below. From the graph, we could determine that correlation way farer than the support vector machine. The value of the RMSE directly gave impact to the prediction outcome.

A final touch is given to the whole research. Based on the findings, Support Vector Machine runs with the data of the other two places. The both data produce few values as below. The run was conducted by applying Nu = 0.6, Capacity = 2.0 and Gamma = 2.0. The model outcome considers the best. As we see the nearest value towards 1 is the best in the support vector machine.

Table 4. Statistical Evaluation total using 8, 12 and 16, V-fold Cross-Validation for Epsilon -Radial basis Function and Nu – Radial basis Function Models

V- Fold		8	12	16
Nu	CC	0.719	0.854	0.84
	MAE	0.00524	0.01309	0.002190
	RMSE	0.84794	0.92412	0.91761
Epsilon	CC	0.843	0.523	0.758
	MAE	0.00748	0.01069	0.00567
	RMSE	0.91815	0.72319	0.87063

Sensitivity analysis in information mining and measurable model building/fitting for the most part alludes to the evaluation of the significance of indicators in the particular (fitted) models. So, given a fitted model with certain model parameters for every indicator, what the impact would be of differing the parameters of the model (for every variable) on the general model fit.

In Statistical Data Miner, affectability examination is accessible through a few alternatives; the specific insights and measures that will be accounted for will subject to the factual or information digging strategy for which the affectability investigation is requested. In Statistical Automated Neural Networks, the program will process the Sums of Squares residuals or misclassification rates for the model when the individual indicator is dispensed with from the neural net; proportions (of the lessened model versus the full model) are likewise reported, and the indicators (in the outcomes table) can be sorted by their significance or pertinence for the specific neural net.

Sensitivity analysis is carried out by using two data mining which are Support vector machine and multilayer prediction. Support vector machine is tested by running with few parameters. Initially, the model is tested by using 3 parameters which are wind speed, humidity and ozone concentration. In the next model, 4 parameters are used which are wind speed, humidity, NOx and ozone concentration. Lastly, 6 parameters which are wind speed, humidity, NOx, So2, No2 and ozone concentration are used to run the model. The values obtain from the run was 0.91214, 0.793095, 0.911043. The patterns of the values increase at 3 parameters then drop when using 4 parameters and increase gradually when using 6 parameters as shown in Table 5.

Besides that, the same model to be run by each parameter is tested one by one. The dependent variable of the run is Ozone Concentration (O_3). The results are tabulated as per below. The results also show that the best value of was 0.889944. These results obtain from the run between humidity hourly average and ozone concentration hourly. The RMSE value of this particular parameter produces better results than another parameter. This could be the one of the main factors that disturbs the reading of ozone concentration. Statistical evaluation total using one to one parameter as shown in Table 6.

Table 5. Statistical Evaluation total using 3, 4 and 6 Parameters

No of SVM	RMSE	Parameters
982	0.91214	3
958	0.793095	4
1238	0.911043	6

Table 6. Statistical Evaluation total using One to One Parameter

Parameters	R	RMSE
Ws hourly avg	0.736	0.857904
Humidity hourly avg	0.792	0.889944
NOx hourly avg	0.587	0.766159
So2 hourly avg	0.189	0.434741
No2 hourly avg	0.277	0.526308
Co hourly avg	0.485	0.696419

4. CONCLUSION

The proposed model has affirmed its capability in predicting ozone concentration with high level of accuracy. Two types of SVM have been investigated in this study. Nu-SVM gives a RMSE value equal to 0.92412 while epsilon- SVM gives a value of 0.91815 which way lesser than Nu- SVM.

ACKNOWLEDGEMENTS

The authors thankfully acknowledge the Ministry of Higher Education Malaysia for the fundamental research grant scheme received grant coded No: FRGS/1/2018/TK10/UNITEN/03/2. The authors would like to appreciate the financial support received from Bold 2025 grant coded RJO 10436494 by Innovation & Research Management Center (iRMC), Universiti Tenaga Nasional, Malaysia

REFERENCES

[1] Abdul-Wahab, S.A., Bouhamra, W., Ettouney, H., et al. (2000). Analysis of Air Pollution at Suhaiba Industrial Area in Kuwait. Toxicological and Environmental Chemistry 78, 213-232.

[2] Abdul-Wahab, S.A., Bakheit, C.S., Al-Alawi, S.M. (2005). Principal Component and Multiple Regression Analysis in Modelling of Ground-level Ozone and Factors Affecting its Concentrations. Environmental Modelling & Software 20, 1263-1271.

[3] Afroz, R., Hassan, M.N., Ibrahim, N.A. (2003). Review of Air Pollution and Health Impacts in Malaysia. Environmental Research 92, p. 71-77. Doi: 10.1016/S0013-9351(02)00059-2.

[4] Afroz, R., Hassan, M.N., Awang, M., Ibrahim, N.A. (2007). Benefits of Air Quality Improvement in Klang Valley Malaysia. Environmental Pollution 30, p. 119-136. Doi: 10.1504/IJEP.2007.014507.

[5] Agirre-Basurko, E., Ibarra-Barastegi, G., Madariaga, I. (2006). Regresson and Multilayer Percepteron-based Models to Forecast Hourly O3 and NO2 Levels in the Bilbao Area. Environmental Modelling & Software 21 (4), 430-446.

[6] Al-Alawi, S.M., Abdul-Wahab, S.A., Bakheit, C.S. (2008). Combining Principal Component Regression and Artificial Neural Networks for more Accurate Predictions of Ground-level Ozone. Environmental Modelling & Software 23, 396-403.

[7] Aljanabi, Q.A., Chik, Z., Allawi, M.F., El-Shafie, A.H., Ahmed, A.N., El-Shafie, A. Support vector regression-based model for prediction of behavior stone column parameters in soft clay under highway embankment. Neural Comput. Appl., 2017, 1–11.

[8] Ali, A., Alfarham, A., Robinson, E., et al. (2008). Tropospheric Ozone Effects on the Productivity of Some Crops in Central Saudi Arabia. Am. J. Environ. Sci. 4, 631-637.

[9] Department of Environment (DoE), Malaysia (2015) MalaysiaEnvironmental Quality Report 2012. Kuala Lumpur: Department of Environment, Ministry of Sciences, Technology and the Environment, Malaysia.

[10] H. Afiq, E. Ahmed, N. Ali, A. Othman, K. Aini, H.M. Mukhlisi, 2013. Daily forecasting of dam water levels: comparing a support vector machine (SVM) model with adaptive neuro fuzzy inference system (ANFIS). Water Resources Management, 27, pp. 3803-3823, 10.1007/s11269-013-0382-4

[11] Intergovernmental Panel on Climate Change (IPCC): Climate Change (2007): The Physical Science Basis, contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., and Miller, H. L., Cambridge Univ. Press, Cambridge, UK and New York, USA.

[12] Tarmizi, A.; Ahmed, A.N.; El-Shafie, A. Dissolved Oxygen Prediction Using Support Vector Machine in Terengganu River. Middle-East J. Sci. Res. 2014, 21, 2182–2188.

[13] Hua, R. (2010). Support Vector Machine Classification and Regression Based Hybrid Modelling Method and Its Application in Raman Spectral Analysis. Chinese Journal of Scientific Instrument 11, 2440-2446.

[14] Garner, M.W., Dorling, S.R. (1998). Artificial Neural Networks, the Multilayer Perceptron. Review of Applications in the Atmospheric Sciences. Atmospheric Environment 32 (14/15), 2627-2636.

[15] Gardner, M.W., Dorling, S.R. (1999). Neural Network Modelling and Prediction of Hourly NOx and NO2 concentrations in Urban Air in London. Atmospheric Environment 33 (5), 709-719.

[16] Hipni, A.; El-shafie, A.; Najah, A.; Karim, OA.; Hussain, A.; Mukhlisin, M. Daily Forecasting of Dam Water Levels: Comparing a Support Vector Machine (SVM) Model With Adaptive Neuro Fuzzy Inference System (ANFIS). Water Resources Management. 2013, Volume 27(10), pp.3803–23.

[17] Floudas, C.; Parlalos, P. Collection of Test Problems for Constrained Global Optimization Algorithms. Springer-Verlag, Lecture Notes in Computation Science, 1990, pp.455

[18] Yan, Jianzhuo.; Zongbao Xu.; Yongchuan Yu.; Hongxia Xu.; Kaili Gao. Application of a Hybrid Optimized BP Network Model to Estimate Water Quality Parameters of Beihai Lake in Beijing. Applied Sciences. 2019, Volume 9(9).

[19] Lai, V.; Najah, A.; Malek, M.A; El-Shafie, A.; Amr El-Shafie, Evolutionary Algorithm For forecasting Mean Sea Level Based on Meta-Heuristic Approach. *International Journal of Civil Engineering and Technology*, 2018, Volume 9(11), pp. 1404- 1413.

[20] Olivia Muslim, T.; Najah, A.; Malek, M.A.; El- Shafie, A.; Amr EL-Shafie. Investigating the Impact of Wind on Sea Level Rise Using Multilayer Perceptron Neural Network (MLP-NN) At Coastal Area, Sabah. *International Journal of Civil Engineering and Technology (IJCIET)* 2018, Volume 9(12), pp. 646–656.

[21] Imani, M.; Kao, H-C.; Lan, W-H.; Kuo, C-Y. Daily sea level prediction at Chiayi coast, Taiwan using extreme learning machine and relevance vector machine. Global and Planetary Change. 2018, Volume 161,pp. 211–221.

[22] Lai, V.; Ahmed, A.N.; Malek, M.; Abdulmohsin Afan, H.; Ibrahim, R.K.; El-Shafie, A.; El-Shafie, A. Modeling the Nonlinearity of Sea Level Oscillations in the Malaysian Coastal Areas Using Machine Learning Algorithms. Sustainability 2019, 11, 4643.

[23] Cherkassky, V.; Xuhui, S.; Mulier, F.M.; Vapnik, V.N. Model complexity control for regression using VC generalization bounds. IEEE Trans. Neural Netw. 1999, 10, 1075–1089.

[24] Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning; Data Mining, Inference and Prediction; Springer: New York, NY, USA, 2001.

[25] Kwok, J.T. Linear dependency between ε and the input noise in ε –support vector regression. In International Conference on Artificial Neural Networks; Dorffner, G., Bishof, H., Hornik, K., Eds.; Springer: Berlin/Heidelberg, Germany, 2001; pp. 405–410.