MAKE WORLD GREEN AGAIN!

Derya ERSOY
21327972
b21327972@cs.hacettepe.edu.tr

Merve Gul KANTARCI
21527104
mervegulkantarci@gmail.com

Ihsan Baran SONMEZ
21427358
ibaranson@gmail.com

Abstract

In this project, we aim to make a research about predicting air pollution level by taking advantage of machine learning algorithms and other works made before. Our dataset consists of hourly meteorological parameters and concentrations of molecular level air pollutants in a certain city. We would like to predict the peak value of the pollutants for the rest of the day by benefitting this dataset. We considered the problem from two different ways. At first, we considered it as a classification problem so we chose using neural network and decision tree to estimate the density level of the pollutants. As the second way, we tried to estimate the peak value so we have picked regression. This was necessary to foresee the AQI (Air Quality Index) of the day because this estimation requires numerical values. For AQI estimation we also tried polynomial regression and classification. For classification we have used support vector machine algorithm. Estimating peak value gave us the chance to estimate the AQI of the day. That is why we consider this approach more important.

1. INTRODUCTION

Air pollution refers to the contamination of the air when harmful particulates and substances are released to the atmosphere in large quantities. It is commonly dense in urban areas due to the unnatural sources like large population, burning fossil fuels, gas release from vehicles, farming chemicals, toxic smoke from factory manufacturing industries and natural sources as dust carried by the wind from locations with very little or no green cover, gases released from the body processes of living beings (Carbon dioxide from humans during respiration, Methane from cattle during digestion, Oxygen from plants during Photosynthesis), smoke from the combustion of various inflammable objects, volcanic eruptions, flora of the area,.

As results of air pollution, we can give many examples. For instance, lead from exhaust fumes attacks the heart, kidneys and nervous system and is particularly damaging to childrens brains. It contaminates soil, urban dust and crops. In the United States, that pollution led to unnaturally high level of lead in the blood of nearly nine out of ten children. From 1973, the US started phasing out leaded gasoline and by 2006 the proportion of children with harmful levels of lead in their blood had dropped to just 1

In Shanghai, according to the Shanghai Daily, approximately 35 percent of days during the first three months of 2013 were considered slightly or very hazardous to health. In October 2013, the World Health Organization (WHO) officially classified air pollution and particulate matter as level Group 1 human carcinogen (cancer-causing), the same level as tobacco.

Plus, founded as a slave labor camp in 1935, the Arctic city of Norilsk is now the worlds largest heavy metals smelting complex and the company responsibleNorilsk Nickelis Russias biggest air polluter. Copper, nickel, lead and other heavy metals taint the soil and water supply while sulfur dioxide emissions contribute to chronic diseases of the lungs, respiratory tracts, and digestive systems, and can result in lung cancer.

Also vehicles travel on a street in Linfen, the most polluted city in China and possibly the world. Hundreds of coal mines, factories and refineries spew out pollutants like sulfur dioxide, coal dust, lead and arsenic into the air and water supply. Three million people choke on coal dust in the evenings, children in the area suffer lead poisoning, while locals have increasing rates of bronchitis, pneumonia and lung cancer.

The AQI is an index for reporting daily air quality. It tells you how clean or polluted your air is, and what associated health effects might be a concern for you. The AQI focuses on health affects you may experience within a few hours or days after breathing polluted air. Different countries have their own air quality indices, corresponding to different national air quality standards. Some of these are the Air Quality Health Index (Canada), the Air Pollution Index (Malaysia), and the Pollutant Standards Index (Singapore), EPA (United States Environmental Protection Agency) .EPA calculates the AQI for five major air pollutants regulated by the Clean Air Act: ground-level ozone, particle pollution (also known as particulate matter), carbon monoxide, sulfur dioxide, and nitrogen dioxide. For each of these pollutants, EPA has established national air quality standards to protect public health .Ground-level ozone and airborne particles are the two pollutants that pose the greatest threat to human health in this country. Think of the AQI as a yardstick that runs from 0 to 500. The higher the AQI value, the greater the level of air pollution and the greater the health concern. For example, an AQI value of 50 represents good air quality with little potential to affect public health, while an AQI value over 300 represents hazardous air quality. An AQI value of 100 generally corresponds to the national air quality standard for the pollutant, which is the level EPA has set to protect public health. AQI values below 100 are generally thought of as satisfactory. When AQI values are above 100, air quality is considered to be unhealthy-at first for certain sensitive groups of people, then for everyone as AQI values get higher.

Increasing AQI is important since air pollution has both acute and chronic effects on human health, affecting a number of different systems and organs. It ranges from minor upper respiratory irritation to chronic respiratory and heart disease, lung cancer, acute respiratory infections in children and chronic bronchitis in adults, aggravating pre-existing heart and lung disease, or asthmatic attacks. In addition, short- and long-term exposures have also been linked with premature mortality and reduced life expectancy.

Prediction of air quality is significant to avoid the risk of being exposed to unclean and hazardous air. This is the point when machine-learning becomes the key solution. Today, the amount of digital data being generated is huge thanks to smart devices and Internet. The current hardware has the capability to reliably store and analyze the massive data and perform massive amount of computations in a reasonable amount of time. This allows to build complex Machine-Learning models with billions of parameters which was not possible a decade ago. This data can be analyzed to make intelligent decisions based on patterns, and Machine Learning helps to do exactly that.

Machine-learning does not provide only a prediction to air quality level, but it also suggests solutions and improvements. This will provide researchers necessary information to work on reducing pollution. The model created can predict the air quality for hypothetical cases For instance, let us assume that there is a new kind of fuel. By adding the gas amount released by fuel to main data, predicting air pollution will show if the fuel is eco-friendly. When a new factory is built in the area, the model can compare the quality level of the air before and after the factory. As a real-time example, for some time intervals of the day, the model can suggest restricting the number of drivers on the road temporarily. While we were searching for a study topic, we came across various projects about this subject. We considered this topic as an up-to-date threat to the environment we live in and found it worth investigating. And this made us interested in this subject and we focused our search on this topic.

2. RELATED WORK

Since the air pollution has a huge effect on the environment and directly affecting human life, there has been a lot of research made about this area to decrease pollution or warning people about hazardous pollution levels. The scientists approach this as a machine-learning problem because forecasting air pollution by using a wide range of data will enable them to check how an improvement will affect air pollution with assumptions and predictions. However, these works still are yet to achieve an absolute solution. That is why there are different approaches with different types of data.

If I compare this work which is basically similar to ours; Firstly, this project has used data obtained form only one station (Athens, Greece) while our data consists of data from many different stations in America. Since the station is fixed they were able to add variables like wind. However, we used more than one station hence we could not use the features that geographical conditions affected directly. Then, this group approached this as a classification problem while we saw it as a regression problem. This caused our work to fall behind of theirs in terms of accuracy. If I widen this subject, they increased the accuracy on their prediction rates by dividing the data into levels with quite big intervals. The lack of harmful examples caused our classification guess to be divided in much smaller intervals which is our projects shortcoming about this subject. This way our prediction was more sensitive but the accuracy was lower. The reference work used MNN as the methods, besides the 6, 7 and 8th hour NO and NO2 values, wind, rain, inversion, solar radiation as features. We used 6, 7 and 8th hour NO2 values, temperature, relative humidity which explained in much more detail at the following parts. (You can access more information under Neural Network title).

For the points the works are different from each other, the reference work completely focused on predicting the peak level of NO2 value while ours has chosen to estimate the AQI value as objective. To achieve this we also needed the peak NO2 value. In addition, we estimated the peak CO value with the same methods. Finally, we have foreseen the maximum AQI value of the pollution according to these peak values within the day.

The past works used neural networks, support vector machine as the algorithm by a majority. Commonly, molecular data has been the main choice of data cate-

gory. Similar to our project, some projects focused on anticipating the peak values of these pollutants. There are even ones that use NO2. On the other hand, some of the projects commonly used ozone or particulate matter (PM) as data domain. In addition, several projects focus on the air quality index (AQI) to classify the level of pollution by government organizations. AQI value is calculated by using relatively various pollutant data types. Also, there are projects uses different algorithms other than we used.
For example, a research made in The University of Auckland, New Zealand about developing an ANN-based air pollution forecasting system. Project basically consists of an ANN-model for predicting NO2 concentration using meteorological values and time variations for a year. Meteorological inputs are wind speed, wind direction, atmospheric temperature, relative humidity and solar radiation. Hour of the day, day of the week, month of the year are the time variations. Data is collected from almost all meteorological stations in New Zealand to ensure the model is applicable on practical use. As method, ANN and Multi linear regression is used. Project concludes that a simple ANN model can give reliable estimations of NO2 concentrations with the given input concepts in this case meteorological data and time variables. Also, it shows that the most effective variable is wind in Auckland, New Zealand. Another work is made in Salamanca, Mexico. This specific project is about a model for estimating PM10 concentrations which is a pollutant not in gas form but particle known as particulate matter- in Salamanca for the next 24 hours. This model uses Multilayer Perceptron Neural Network and clustering algorithm. The dataset consists of historical meteorological variables and concentrations of PM10 collected from three different stations. Clustering algorithms which are Kmeans and Fuzzy CMeans are for finding relationship between PM10 and meteorological variables. These links will help to get the extra information for the prediction process. The model is compared with a Multilayer Perceptron and multiple linear regression. The results show that ANNs combined with clustering algorithms has better generalization capacities than those based on a simple ANN and multiple linear regression.

In addition, there is a system that is worked on to forecast daily AQI (Air Quality Index) in Macau using neural network. This is a project to predict one day ahead daily air quality index in the area which is data gathered in Macau. Various input selection are tested against different number of hidden layer neurons. It is found experimentally that using only the past three days values as input with 8 neurons in the hidden layer give the best testing results. Data is from April to June 1999 and it is compared to July 1999. Initial assessment of the ANN application on air quality forecasting in Macau is performed with satisfactory results. This shows that an efficient air quality prediction system is very likely possible in Macau by applying neural network but it can still be improved.

These are the approaches we mostly inspected and took as references. However, there are many other works in literature. The options are limitless because there is still data and methods to be tested. As air pollution keeps growing, the works will keep increasing.

3. THE APPROACH

Our goal was to predict the AQI value at the time that air pollution is the highest in a day. In order to succeed, instead of using the all of the hourly measurements in a day, we have benefitted from a related work -explained in detail in related work section- and it guided us to decide on using early concentrations of pollutants and meteorological parameters. We have tried to estimate the AQI of the day directly from these features, but we got better results when the peak values of the pollutants were estimated firstly and the AQI model was fed by these estimations. In addition to this improvement, we have considered peak value estimation useful since this was also a research topic. As we have seen from previous works, predicting the peak value of a pollutant can be considered as a classification problem or it might require regression analysis. We examined these two approaches as well. The regression analysis gave us to opportunity to estimate the AQI level, but we didnt implement a model for AQI-prediction using classification results because of the reasons mentioned in section 3.2.1. In general, it can be said that there are two main branches of this research: discrete case and continuous case.

3.1. Dataset

In order to build our dataset for our project, we first made researches using related works. These researches gave us an idea on features to be used. Thus, we have started to collect our data by searching on internet. We came across a very broad archive including datasets for hourly and daily measurements for a variety of air quality parameters. There were lots of useless information in these files, so that we needed to eliminate them and simplify the set. After we made this process for each of the features (NO2, CO, relative humidity, temperature), the values at 6, 7, 8 hours are kept and the maximum value of the pollutant density is added as the label. These process is done for NO2 and CO since they are the base pollutants we have used in our project. For temperature and relative humidity only the 3 hourly measurements were required. Another step for finalizing our dataset was combining this features into a single file since the features were originally kept in separate files. The resulting set was a matrix with dimensions 12000x22. Using a single npy file as the dataset made easier later implementations. The below shows a row of this matrix. While testing our implementations, we didnt prefer to choose cross-validation method since our train set and test set includes enough samples. Our train set includes 10006 samples and the test set includes 1237 samples. While dividing data into train and test sets we have shuffled the data in order to distribute the different samples uniformly.

3.2. The Classification Approach

Even though we continued with regression models, we also examined the classification methods. We saw that most of the previous related works used multi-layer neural network. This encouraged us to start with neural network. We have made experiments on activation function, layer size, neuron size in layers. After the tests made, we got the best accuracy result with a network 3 hidden layers. Layers consisted of 22, 12 and 6 neurons respectively. After these values, we did not observe a positive effect on accuracy when neuron size is increased more, however time complexity increased much. We also tried to increase the layer size, but similarly this also did not make a significant change. We got the best results using logistic function as the activation function. We did not get as good results as we expected with relu and tanh functions.
Another model is created by implementing decision tree algorithm for the classification case. For true values, we have used the labels that are already produced in previous parts. Since there are not many hyperparameters in this approach, it did not take much effort to reach an optimum model. We gave priority avoid overfitting, for achieving this, we have worked on finding an optimum max-depth scaler.

3.2.1 AQI Estimation

When we approached AQI-prediction problem with classification methods, the obtained results were not suitable and sensible enough. In the case that peak values are classified by some levels, since the levels have a wide range for peak values, using obtained results for AQI estimation would be misleading. In detail, estimating AQI value requires 2 features in our project: the peak value of concentrations of NO2 and CO of the day. These two attributes are classified into 4, 6 or 12 levels. We did not find it efficient to estimate a scalar value with these inputs. 12-level classification might be useful but the accuracy for that model was relatively low. Besides, the results as levels are considered enough for previous related works. Because of these reasons, we thought that it might be more meaningful leaving this part as it was.

3.3. The Regression Approach

Peak value estimation seemed us as a regression problem since it is a continuous case originally. We have worked on linear regression, polynomial regression and support-vector regressor.
Firstly, we started our experiments with linear regression and we have tried to estimate peak value of NO2 in a day. The results were promising even at this first try but we were not able to make any improvements on this model. L2 norm and l1 norm is tried for normalization, standardization is tried as well but none of them made a progress on results. Since this linear regression model did not provide good results, we have worked on kernel ridge regression with linear kernel function. This approach resulted worse.
The next approach we have worked on was polynomial regression algorithm. For this model, we made improvements by making experiments on degree of function. Even though the model was not very tended to overfit, we made experiments on regularization constant and tried to find an optimum one. In this model, normalization made a huge difference and l2 norm is used for normalization.
Another approach was using support-vector regressor, but the results were not worth evaluating since this model resulted the worst accuracy in this part. We have ended up deciding on 5-order polynomial regression with 0.001 as the regularization constant.

3.3.1 AQI Estimation

In the traditional way, AQI is either a scalar which means the continuous case- or the index table is divided into 5 levels (The index table can be seen from introduction section). Apart from the traditional table which points to a new level in each 50 or 100 units, we have divided our index table into 7 levels. The wide of each level is 15 units. We have followed such a way because our data was not including samples for extreme cases that corresponds to very polluted air. Instead of continuing with the samples that map to first two levels of the table, we thought of this approach, and considered this as more sensible. Thus, the minimum AQI belongs to level 1 and maximum AQI belongs to level 7 in our dataset.
We have implemented two models for the cases above, as in the peak value estimation part: classification and regression models. The model used for AQI estimation takes the pollutant densities -peak values of NO2 and CO in our project- as input. It outputs the predicted AQI of the day in the regression model and it outputs the predicted level of AQI of the day in the classification model.
Firstly, we tested this part with a dataset consists of peak values of NO2, peak values of CO and AQI of days to see the possible success at estimating AQI value. The scores were higher than 0.95 for both regression and classification

part. This supported our idea. However, we did not take the error in the input data into consideration. When we fed the models with estimations from the first phase, the peak value estimation, the results were not as high as the before. The error in first phase was directly misleading the output of this part. In the second approach, instead of using this model trained with true values, we have used a train set consist of the estimated peak values with the models mentioned in section 4.3. The aim was to train the AQI model for this kind of errors. This gave us a bit more accurate results in classification method.
For the classification model, we have implemented support-vector machine algorithm. We made experiments on kernel function, and the results are examined. We obtained the best result with linear kernel. For the regression analysis, we have used polynomial regression and kernel ridge regression both resulted similar, and we decided on polynomial regression, since we find it more consistent with different test sets.

4. THE EXPERIMENTAL RESULTS

While we were deciding on hyperparameters and models to be used, we benefitted from our experiments. After making comparisons, we have eliminated some of the models. In this section, we discussed this process by showing our steps.
The scores for regression type models are calculated usingr^2 score and as the error function mean-squared error is used.

4.1. Peak Value Level Estimation

As we mentioned before, there are many hyperparameters that can be changed in multilayer neural network implementation. Especially neuron size and layer size are strongly correlated to each other. As an example from this project, for 4 and 6 levels prediction, we got the same results while using the network consist of 2 hidden layers with 15, 8 neurons respectively and while using the network consists of 3 hidden layers with 22, 12, 6 neurons. For 12-level prediction, we got higher results with the networks consisting more layers. In short, while increasing hidden layer size did not affect the scores in 6-level set, we saw a significant change in 12-level set between 2 and 3 hidden layers. As the activation function we have tried on tanh, logistic and relu functions. Relu and logistic functions resulted similar scores in every level system. However, tanh function resulted the worst scores in all 3 level systems. Thats why we have decided on logistic function as the activation function.

In decision tree model, there are not many parameters that might affect the results. We only made experiments on maximum depth. Until reaching an optimum value -which is decided as 6 in our project-, the scores first increases then decreases. The features cause the increase at first. But after some point increasing maximum depth causes overfit and the model memorizes the train set. This makes the model tended to make errors in unusual samples. The experiments held while deciding on maximum depth are shown below.

Table 1. Activation Functions' Scores

Activation Functions	4 labels	6 labels	12 labels
Tanh	0,85	0,78	0,57
Logictic	0,87	0,79	0,58
ReLU	0,86	0,78	0,58

Table 2. Effect of Depth to the Score

Maximum Depth	4 labels	6 labels	12 labels
3	0,77	0,56	0,85
4	0,78	0,58	0,88
5	0,78	0,60	0,86
6	0,80	0,62	0,87
7	0,80	0,61	0,87
8	0,78	0,61	0,87
9	0,77	0,59	0,87
200	0,70	0,51	0,83

4.2. Peak Value Estimation

We have worked on standardizing and normalizing the data but none of them made any improvement on results. We have obtained a score of 0,78 for CO estimation and a score of 0,74 for NO2 estimation. We also implemented linear kernel ridge regression but this way the score decreased.

Table 3. r^2 scores with 3 order in polynomial regression

Pollutants/Scores	Degree=3			
Reg Const	0,5	0,1	0,001	1e-5
NO2	0,74	0,76	0,78	0,79
CO	0,78	0,79	0,80	0,80

Table 4. r^2 scores with 5 order in polynomial regression

Pollutants/Scores	Degree=5			
Reg Const	0,5	0,1	0,001	1e-5
NO2	0,77	0,77	0,79	0,79
CO	0,79	0,79	0,80	0,80

Table 3 and 4 above shows the experiments made on regularization constant and function degree. There was not a big difference between 3-order and 5-order when the regularization constant picked wisely. However, we preferred to choose 5-order polynomial regression since generally it outputs better results.

4.3. AQI Estimation

Table 5. Scores with Degrees

Degree	3	5	7
Scores	0,79	0,79	0,78

Table 6. Scores with different constants

Reg Const	1	0,5	0,1	0,01	0,001	1e-5
Scores	0,71	0,74	0,78	0,78	0,79	0,79

For the continuous estimation of AQI, we have made a series of experiments on linear kernel ridge regression model and polynomial model. There was not much to change for linear kernel ridge regression, but we made some experiment on polynomial regression. Our dataset was not tended to overfit with thats why very small regularization constants did not make significant changes. We have avoided using very large alpha values since it caused to underfit. After the experiments, alpha=0.001 seemed a good solution. While deciding on degree size of the function, we have observed that larger degree could cause a huge increase in time complexity (i.e. 10-order), but we did not need such a high degree size, even the 7-order polynomial regression did not fit better to our data. We had to decide between 3 and 5 as degree size, when we made experiments in different test sets, we concluded that 5-order polynomial regression was likely to be a better solution since in most cases it fit better. At the end, we compared the linear krr and polynomial regression, we have observed that their scores are close to each other (both 0.79). At the end, we decided on 5-order polynomial regression with 0.001 as the regularization constant since this way we obtained the best results in a few different test sets.

Table 7. SVM Model with different kernel

Kernel Function	Sigmoid	Rbf	Linear
Scores	0,42	0,75	0,76

For the predicting AQI-level according to our level system we have mentioned above, we have used SVM approach. SVM was very successful when we made an experiment using true peak values of pollutants. It gives a score of 1.00 on estimating AQI-level and a score of 0.96 on estimating AQI. But when the model tested with the output from first phase, because of the error in the estimated peak values, the accuracy decreased drastically. This was expected since we couldnt prevent the error in the first part. We saw that the most accurate result was obtained using linear kernel function. Table 7 shows this comparison.

As its seen from the plot above, the best accuracy result gets close to the worst case of the first phase. The source of the mispredictions here is the error in the input data. Since the input of this part is the output of the first phase, the error in the first phase directly misleads the estimations. In order to train the model for this kind of errors, we used a train set consists of the estimated peak values instead of a train set consists of true peak values. The progress made in SVM model with this idea is shown in the table below. The Polynomial model didnt make progress with help of this idea. Table 8 shows the difference.

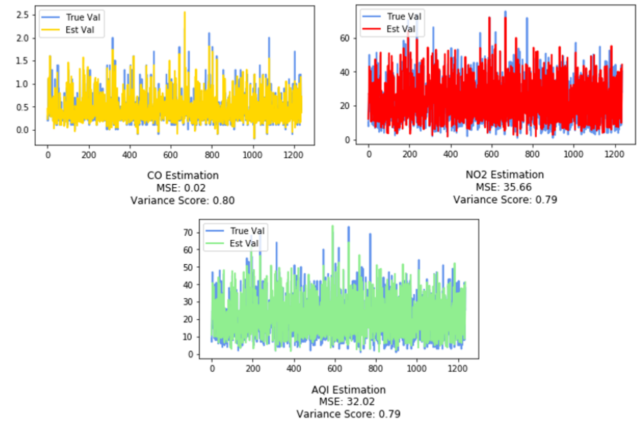

Figure 1. Test scores and errors with polynomial regression for peak value estimation and AQI estimation

Table 8. Scores with different input data

MODEL	Scores using true values for data	Scores using est values from first phase
Poly Reg	0,79	0,79
Linear KRR	0,78	0,79
SVM	0,74	0,76

5. CONCLUSION

In this project we have tested many approaches about real-time air prediction which is our goal. We can categorize these approaches under two titles which are regression and classification. If we divide the problem into two stages, we needed to predict the maximum values of CO and NO2 molecules within the day as first stage. We have tried classification methods as MNN method which is commonly used in previous works and our observations showed us that MNN is the classification method which has given the best score. However; for the second stages input, we needed continuous values instead of labels. To achieve this, we decided to use regression. For the second stage, using the peak values of CO and NO2 molecules, we have tried to estimate the maximum air pollution within the day. In this stage we have obtained results by using both regression and classification. As we closely inspect the results in the drawn charts, we can easily say that the system which we set up on extreme minimum and extreme maximum values does not give the required result. At this point, besides not being able to find enough elements for these groups effecting the most, the extra increase on the amount of CO and NO2 that comes from the unpredictable events within the day caused the results to be misleading. If we exemplify this, we can give the unexpected fires and construction works. These affects our results directly but since they did not affect the early hour measurements, they raised incorrect results. Some of the factors that we have taken no account of like the geographical structure, state of wind is also a factor that is able to affect the measurements directly at the late hours within the day. We have worked on multiple stations not one region and not found objective wind measurement hence we could not used these features in our system. There are many factors that affect air pollution but we solely used the ones have the most effect. As an instance, we observed that traffic density has an effect on air pollution in the work MIT did about air pollution estimation. However, we could not use the traffic density because we could not find traffic data which can support our used data.

About the improvements can be done to use this system in real life, we can say that our predictions success rate will increase if we, for example, update the learnt dataset through determined needs in time or provide a different dataset for each station or do not add the days that unpredictable events happened.

The obtained results upholds the usage of machine-learning methods on air pollution forecasting. This way, without a dense observation network but with less cost learning system, we can achieve successful guesses with high accuracy about air quality.

6. REFERENCES

1 - Madhavi Anushka Elangasinghe, Naresh Singhal, Kim N. Dirks, Jennifer A. Salmond. Development Of An ANNBased Air Pollution Forecasting System With Explicit Knowledge Through Sensitivity Analysis. Atmospheric Pollution Research Vol.5 696-708. 2014.

2 - Maria Guadalupe Cortina-Januchs, Joel Quintanilla-Dominguez, Antonio Vega-Corona, Diego Andina. Development Of A Model For Forecasting Of PM10 Concentrations In Salamanca, Mexico. Atmospheric Pollution Research Vol.6. 626-634. 2015

3 - K. M. Mok, S. C. Tarn, P. Van , L. H. A Neural Network Forecasting System For Daily Air Quality Index In Macau. Air Pollution VIII, C.A. Brebbia, H. Power, J.W.S Longhurst (Editors). 2000

4 - Elias Kalapanidas, Nikolaos Avouris. Applying Machine Learning Techniques In Air Quality Prediction. University of Patras. September 1999

5 - Junsub Yi, Victor R. Prybutok. A Neural Network Model Forecasting For Prediction Of Daily Maximum Ozone Concentration In An Industrialized Urban Area. Environmental Pollution Vol.92 No.3 Pp. 349-357. 1996

6 - Mario Catalanoa, Fabio Galatioto, Margaret Bell, Anil Namdeo, Angela S. Bergantino. Improving the prediction of air pollution peak episodes generated by urban transport networks. Environmental Science ,Policy Vol.60. 6983. 2016.

7 - Hao Wu, Stefan Reis, Chun Lin, Mathew R. Heal. Effect of monitoring network design on land use regression models for estimating residential NO2 concentration. Atmospheric Environment Vol.149. 24-33. 2017.

8 - Xiao Feng, Qi Li, Yajie Zhu, Junxiong Hou, Lingyan Jin, Jingjie Wang. Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. Atmospheric Environment. Vol 107. 118-128. 2015

9 - Akbar Rahimi. Short-term prediction of NO2 and NOx concentrations using multilayer perceptron neural network: a case study of Tabriz, Iran. Ecological Processes 6:4. 2017.

10 - Stavros Perantonis, Georgios Amanatidis, S.J. Varoufakis, J.G. Bartzis. Neural Network Techniques for SO2 Episode Prediction. National Center for Scientific Research. 153 10. January 1994.

11 - Dimitrios Melas , Ioannis Kioutsioukis, Ioannis C. Ziomas. Neural Network Model for Predicting Peak Photochemical Pollutant Levels. Journal of the Air, Waste Management Association (1995). May 2000