Identification of outliers in pollution concentration levels using anomaly detection

T.R.V.Anandharajan
Associate Professor
Velammal Institute Of Technology
trvanandharajan@gmail.com

K.k.Vignajeth
Graduate Student
Velammal Institute Of Technology
vignajeth@gmail.com

G.Abhishek Hariharan
Graduate Student
Velammal Institute Of Technology
abhishekhariharan1995@gmail.com

R.Jijendiran
Graduate Student
Velammal Institute Of Technology
jijendiranravichandran@gmail.com

Abstract— **Anomaly detection is generally an identification of any odd or anomalous data sometimes even called as an outlier from a give pattern of data. It involves machine learning technique to learn the data and determine the outliers based on a probability condition. Machine learning, a branch of artificial intelligence plays a vital role in analyzing the data and identifies the outliers with a good probability. The objective of this paper is to determine the outlier of pollutant's concentration based on anomaly detection techniques and describe the air quality standards of the particular area**.

IndexTerms: AirPollution, MachineLearning, Anomaly detection, Air quality Index

I.INTRODUCTION

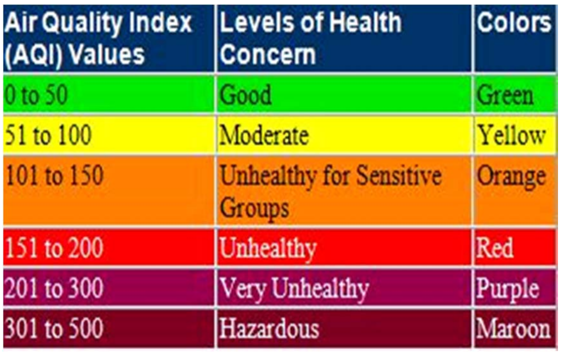
Pollution is a major concern in today's world as we mostly depend on fossil fuels for all purposes. This has caused various problems in the environment leading to issues like global warming and major health hazards. Since energy production by eco friendly methods are not equivalent to energy released by fossil fuels, the dependence on fossil fuels remains the same. The pollution level gradually increases year by year and hence we have to adapt accordingly. In order to adapt, it becomes necessary for the people to get alerted during high pollution concentration levels so that they can deploy preventive measures to protect themselves. The high concentration levels which are hazardous to people are indicated as outliers which are estimated by anomaly detection technique. Anomaly detection is a part of machine learning which comes under unsupervised learning. Unsupervised learning is a technique by which the machine learns from unlabeled data sets and produces a suitable outcome. Unsupervised learning also relies on density estimation methods which help us to find the anomalous data. The concept of detecting the anomalous data by probability density estimation methods is followed here.

II.PERFORMANCE PARAMETERS

A) AQI (Air quality index)

AQI is a number or an index used by governments to measure the quality of air. A high AQI index assures that the air (atmosphere) is polluted and explains the adverse effects of this high concentration level which causes serious health effects. These standards vary according to regions in our planet. Different countries have different levels or indices which vary according to their geographical locations.

This paper uses the AQI referenced from the website:(www3.epa.gov)



Air Quality Index (AQI) Values	Levels of Health Concern	Colors
0 to 50	Good	Green
51 to 100	Moderate	Yellow
101 to 150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

Fig.1.1 The AQI Table [5]

This table is well suited only for United states of America .Based on the air quality index the quality of air is determined and levels of health concern are tabulated

978-1-5090-0082-1/16/$31.00 ©2016 IEEE

correspondingly. The system developed in this paper has the ability to denote AQI greater than 101 of any gas that are harmful.

B) AIR QUALITY INDEX IN NEW YORK/NORTHERN NEW JERSEY-LONG ISLAND:

We have taken the air quality index in a particular region for the estimation of pollutants concentration level .The data has been taken for the year 2014. The following pollutants are taken into consideration: 1.Carbon monoxide 2. Nitrogen- di-oxide 3. Sulphur- di- oxide 4. Particulate matter (PM-2.5) 5. Particulate matter (PM-10) and 6. Ozone

The measurements used in determining the Air quality index are given below:

$$\text{AQI} = \frac{\text{pollution } Data\ Reading}{Standard} *100 \qquad (1)$$

As the conversion of raw data readings were converted into AQI by the EPA (Environmental protection Agency U.S.A) it became easy to proceed with the analysis

The units used by the EPA in measuring various gas concentrations levels are shown below:

Pollutant	**Units used for air quality data**
Ozone	pphm (parts per hundred million)
Nitrogen dioxide	pphm (parts per hundred million)
Carbon monoxide	ppm (parts per million)
Sulfur dioxide	pphm (parts per hundred million)
Particles or Particulate matter	µg/m^3 (micrograms per cubic meter)

Fig.1.2 Pollutants Units

III.FUNCTIONS

A.MEAN:

Mean (μ_j) is defined as the average of the dataset $x_j^{(i)}$.

$$\mu_j = \frac{1}{m} * \sum_{i=1}^{m} x_j^{(i)} \qquad (2)$$

Where m is the total number of rows in the matrix. The variables i and j are used to point the elements in the matrix where j denotes the number of columns.

B.VARIANCE**:**

Variance (σ_j) represents the range of gas concentration level for a particular gas. The variance is given by:

$$\sigma_j^2 = \frac{1}{m} * \sum_{i=1}^{m} \left(x_j^i - \mu_j\right)^2 \qquad (3)$$

A zero variance indicates that all points are identical. Variance is always non negative.

C.GAUSSIAN DISTRIBUTION**:**

The probability density function of Gaussian distribution or normal distribution is very important to determine the outliers present in the data. In a Gaussian distribution the general formula to find the probability density function is given as:

$$P(x) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_{\{j\}}} * \exp\left(\frac{-\left(x_{\{j\}} - \mu_{\{j\}}\right)^2}{2\sigma_{\{j\}}^{\{2\}}}\right) \qquad (4)$$

Where n is given by the total number of columns (the number of gases).for multivariate Gaussian distribution the total probability is equal to the product of the probability of each gases.

D. FI SCORE**:**

The F_1 score is generally a measure of a test's or an outcome's accuracy. It is dependent on both the precision and the recall of the test to compute the F1 score.

E. PRESICION:

$$\text{Precision} = \frac{\text{True postitives}}{\text{Number of } predicted \text{ positives}} \qquad (5)$$

Number of predicted Positives=True positives + False positives.

True positives – The training algorithm predicts that it is positive when the actual class is also positive.

False positives - The training algorithm predicts that it is positive when the actual class is negative.

F .RECALL:

$$\text{Recall} = \frac{\text{True positives}}{\text{Number of actual positives}} \qquad (6)$$

Number of actual positives=True positives + False negatives.

False negatives - The training algorithm predicts that it is negative when the actual class is positive.

And hence the F1 score is given as

$$\text{F1 score}=\frac{2*precision*recall}{precision+recall} \qquad (7)$$

It is very important for the training set to determine the correct anomalies.

IV. METHODOLOGY

The one year data of 2014 of different gas concentrations that cause pollution are collected and partitioned in the three following data set. They are Training set, Cross Validation set and the Test set. Sixty percent (60%) of data is set for the training set, while twenty percent (20%) of the data is set for the Cross validation set and remaining 20% is allocated for the Test set.

The training set is used to find the probability of gases using Gaussian distribution for the reading of each day. The Cross validation set is used to determine the F1 score to make sure that the system is efficient and accurate in determining the anomalous data (outliers) by finding the best value of epsilon. The value epsilon is a probability value which is found by probability density function and is found to be in the range of 0 to 1. This act as a threshold to all the gas's probability .Here gases with lower probability value when compared with the value of threshold (epsilon) are said to be anomalous values or the outliers.

The test set is used in finding out how well the system is perfect in detecting the anomalous data. A good value of accuracy obtained from the test set will show how well the system is functioning well in determining the anomalous values.The entire system is built with the software MATLAB. Hence the data are processed in matrix formats. The training set, Cross validation set and test set are taken in a matrix and are processed to find out the probability density value using Gaussian distribution. The matrix is taken with data of two gases, each gas in a column. The mean, variance and multivariate Gaussian distribution as represented by (2), (3) and (4) are applied.

The multivariate Gaussian determines a probability value which is the product of the probability of gases is compared with the threshold obtained using the F1 score and the anomalies are detected.

In order to calculate the F1 score a separate labeled matrix of 0's and 1's is taken for the cross validation data set which is also called as ground truth table. All the values representing 1's are the anomalous data in the cross validation set and zeros are (non anomalous) data. The data (concentrations levels) represented by 0's will not affect the environment. The F1 score deals with two important parameters called Precision and Recall.

Of all the anomalies found what fraction of the data actually is an anomaly is called Precision. Of all the actual anomalies what fraction did we accurately detect it as anomaly is called Recall.

The precision and recall uses true positives, false positives and false negatives as given by equations (5) and (6). The threshold value (epsilon) which is a probability must be in between the maximum and minimum probability range of the gases and the probability less than epsilon are marked as anomalous. In order to find the best epsilon value, F1 score is used. A binary matrix is generated based on the condition when the prediction where the probability value is lesser than epsilon. A ground truth table consisting of combinations of 0's and 1's of both predictions and actual values is used in finding true positives, false positives and false negatives.

The plots show only combination of two pollutants at a time. Since pollution depends on various pollutants, we have to take all the pollutants into account. It's not possible to show the plots for more than 2 pollutants at a time and hence only the outliers and their number have been determined. And finally a plot showing the combination of two pollutants along with the probability density value is plotted and the detected anomalies are spotted differentiating the normal and abnormal (outliers) data.

V.GAUSSIAN DISTRIBUTION PLOT:

The plots (Fig 1.3,1.4,1.5,1.6,1.7,1.8) plot the histogram of data and probability density(Gaussian distribution) line. It is required to have the histogram of data and its Gaussian distribution. The Gaussian description form can be obtained by either taking logarithm of the data or using higher order powers. These manipulations do not change the probability they in turn improve the detection of anomalies. The redline in the plots give the normal or Gaussian distribution probability curve of the data. The green box like shapes denote the data in the plot.

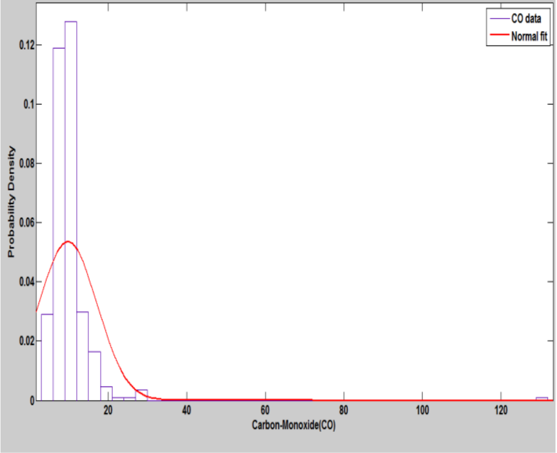


Fig.1.3 Gaussian distribution of CO

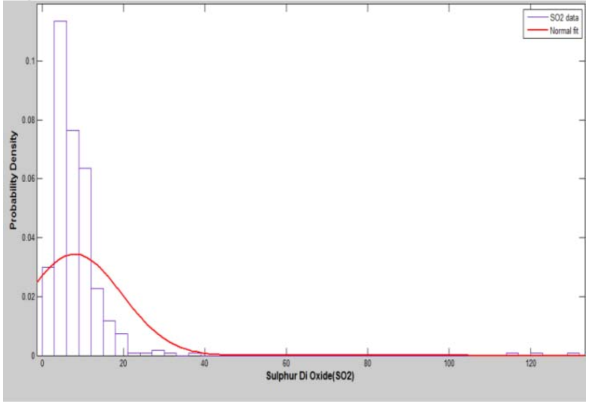


Fig.1.4 Gaussian distribution of SO2

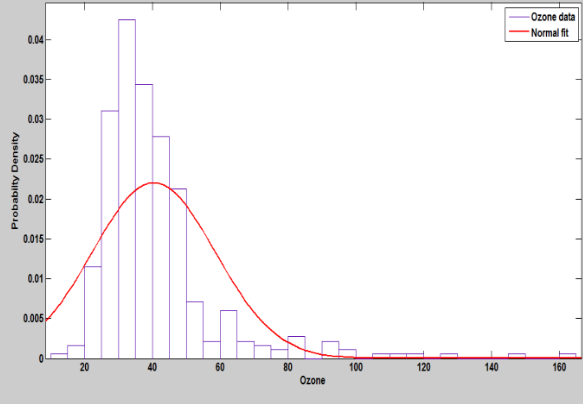


Fig.1.5 Gaussian distribution of Ozone

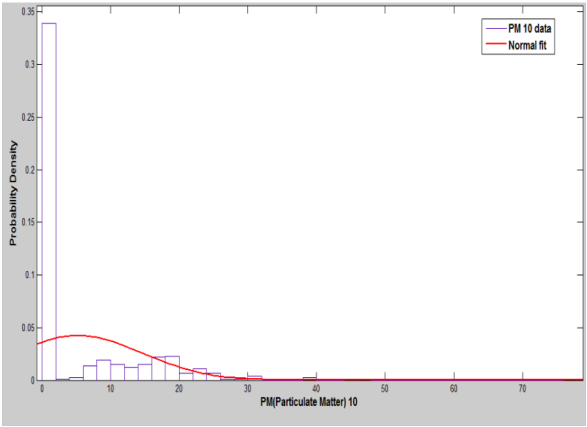


Fig.1.6 Gaussian distribution of PM10

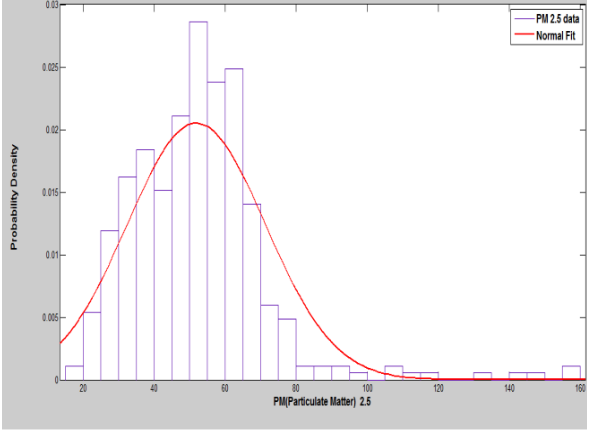


Fig 1.7 Gaussian distribution of PM2.5

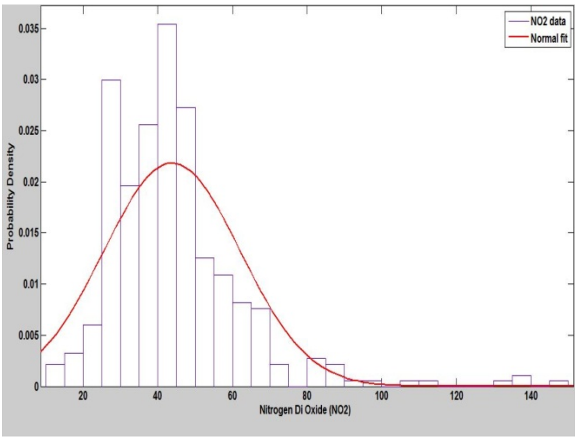


Fig.1.8. Gaussian distribution of NO_2

VI. PLOTS

The visualization of the detected outliers is shown with the help of matlab plots. The following figures show the combination of two gases along with their probability density levels. There are 3 plots, Carbon monoxide vs. Ozone, Nitrogen dioxide vs. Sulphur dioxide and Particulate matter (PM) 10 vs. PM 2.5. Among them there are two dimensional and three dimensional plots. In these plots red colored square frames denote the outliers. In the 2-D plots (Fig 1.9, 1.13, 1.11) the contour circles represent the normal data. The one with the highest epsilon value (probability density) are in red while blue is the least and the colors vary depending on the descending values of probability density.

In the 3-D plots (Fig 1.12, 1.10, 1.14) the pyramid like structure represents the normal data's probabilities and the data with high epsilon value are indicated in red. It is similar to the 2-D plot except that it has another axis namely the probability density axis. The AQI's as explained earlier are taken along the two axes depending upon the gas combination and their probability density values on the other axis (for 3-D plots).As there was no high level of pollution concentration in the New York/Northern new Jersey /Long Island region. some anomalous data were added in the original data to see how well the system is efficient in detecting the anomalous data.

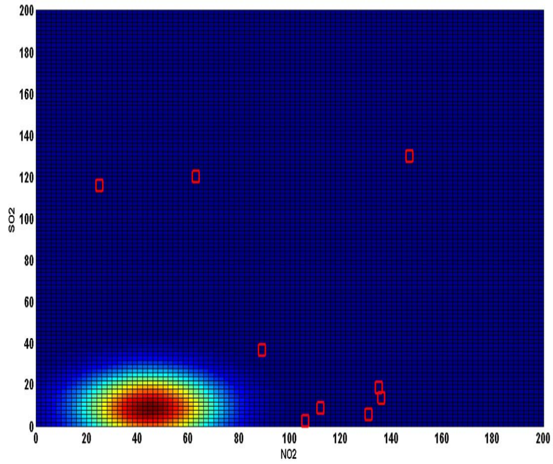


Fig1.9 Nitrogen Di Oxide Vs Sulphur Di Oxide in 2D

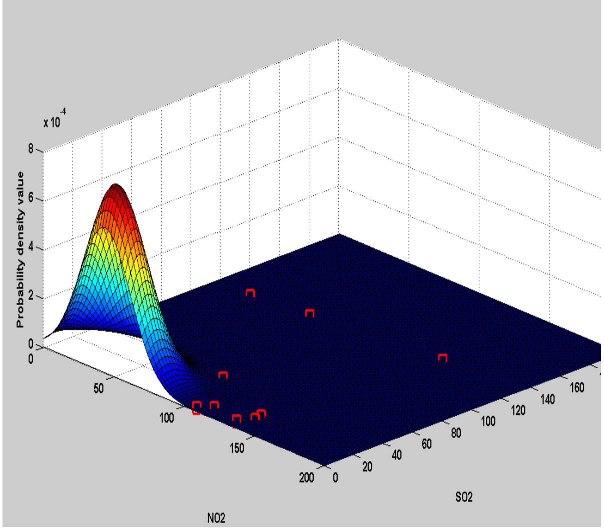


Fig.1.10 Nitrogen Di Oxide Vs Sulphur Di Oxide in 3D

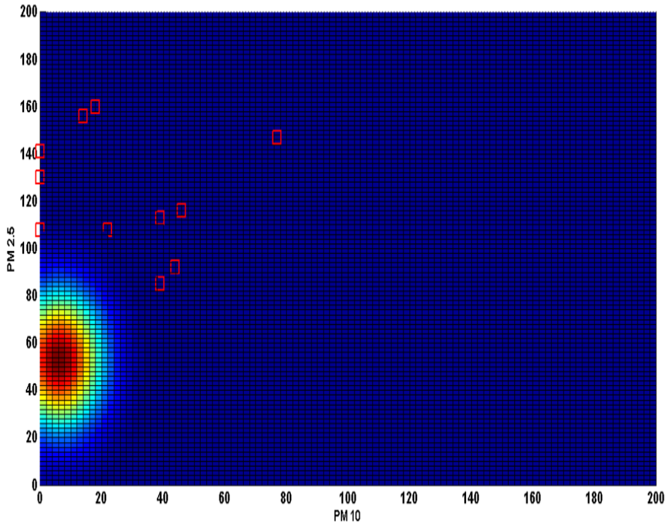


Fig.1.11 PM 10 Vs PM2.5 (Particulate Matter)in 2D

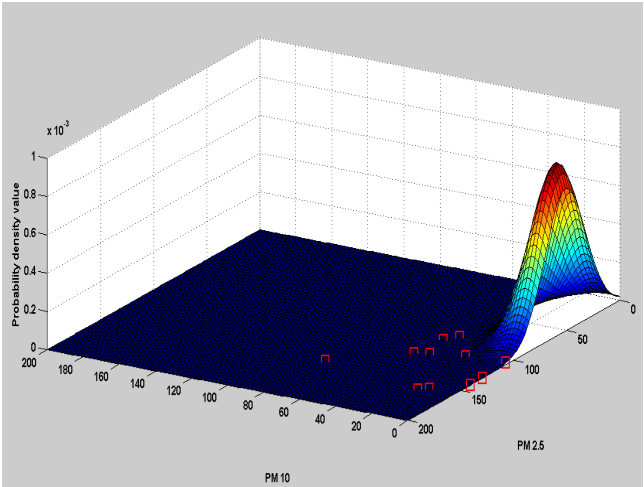


Fig 1.12 PM 10 Vs PM2.5 (Particulate Matter) in 3D

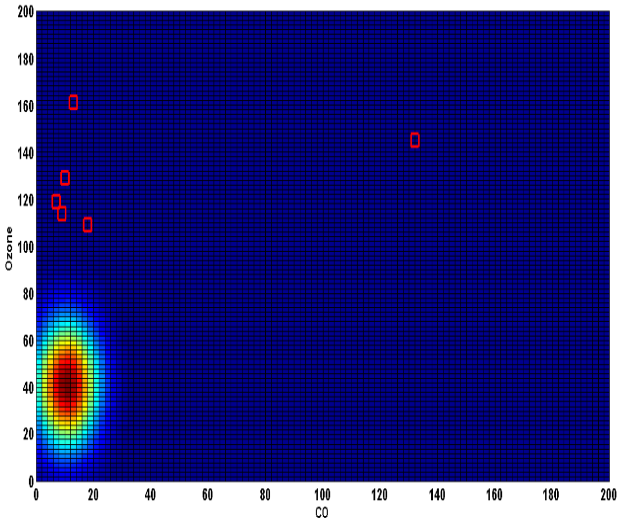


Fig.1.13 Carbon Mon oxide Vs Ozone in 2D

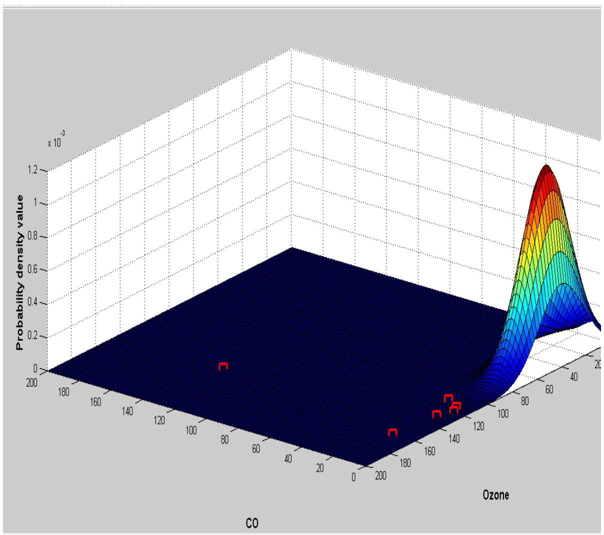


Fig.1.14 Carbon Monoxide Vs ozone in 3D

VII.RESULT

All the Gases (fig 1.2) that are hazardous to the environment were taken and multivariate Gaussian distribution was implemented. An anomalous data(test data) is fed into the system in order to check whether the system is successful in determining the Anomalous data .

The anomalous data sent was [80 4 96 10] which in the form of a matrix with each column denoting the gases –Ozone, Sulphur-Di-Oxide, Nitrogen Di Oxide and Carbon Monoxide respectively.

On training the system, it produced a threshold(epsilon) of 0.0006*1.0e-007 and the cross validation set epsilon (probability) was 3.643036e-008.As the epsilon value of the test data set was lesser than the cross validation epsilon, the sample or test data was accurately identified as an anomaly.

VIII. CONCLUSION

The undesirable levels of the pollutants are detected such that it would provide vital information to the civilians living in the designated area. In a nutshell, the outliers help in providing the information of the anomalous (danger) levels of pollutants in an area. . So that people may consider the need to take control over it as soon as possible. Based on the dangerous levels of several gases, one could easily identify the source for the pollutant and take remedial measures. People living in areas where their lives are more prone to hazardous diseases can be alerted and avoid such happenings. These pollutants cause respiratory problems and sometimes due to prolonged exposure they become carcinogenic.

This paper not only provides the anomalous levels of pollutants but also an easier way of describing it using catchy visual plots. The major drawback is that anomaly detection algorithm suits well only when there are few anomalous data. When there are more data which are not normal, logistic regression by classification methods are the best suited ones. In areas where the atmosphere is constantly polluted with higher AQI values, the effectiveness of anomaly detection algorithm is reduced. In future this algorithm can be made more flexible and developed into a module which can be used by the common man in an efficient and transparent manner.

IX. ACKNOWLEDGEMENT

The idea and the implementation of this paper would not have been successful without the help of our supporting professor Dr.T.R.V.Anandharajan. The transparency in acquiring data which made the paper to obtain results could not have been accomplished without the help of the EPA (United States Environmental Protection Agency).

X. REFERENCE

[1] Sudheendra,Hangal,Monika S,.Lam.:Tracking down software bugs using Automatic Anomaly detection. Proceedings of the 24th International Conference on Software Engineering Pages 291-301.

[2] P. Garcı´a-Teodoro,,J. Dı´az-Verdejo,G. Macia´-Ferna´ndez,E.Va´zquez.:Anomaly-based network intrusion detection: Techniques, systems and challenges. Computers & Security 28 (2009) Page 18–28.

[3] Richard O. Lane ,David.,A. Nevell ,Steven D. Hayward,Thomas W. Beaney.: Maritime anomaly detection and threat assessment. Information Fusion (FUSION), 2010 13th Conference on 26-29 July 2010 Pages: 1 – 8 ISBN: 978-0-9824438-1-1

[4] Daniel W.Urish.:The Practical Application of Surface Electrical Resistivity to Detection of Ground- Water Pollution. National Ground Water Association Volume 22, Issue 3 May 1984 Pages 342–343

[5] https://en.wikipedia.org/wiki/Air_quality_index