International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 11 No. 2 Nov. 2014, pp. 237-247
© 2014 Innovative Space of Scientific Research Journals
http://www.ijisr.issr-journals.org/

DATA MINING INDUSTRIAL AIR POLLUTION DATA FOR TREND ANALYSIS AND AIR QUALITY INDEX ASSESSMENT USING A NOVEL BACK-END AQMS APPLICATION SOFTWARE

E. O. Ofoegbu[1], M. A. Fayemiwo[2], and M. O. Omisore[2]

[1]Department of Computer Engineering,
Oduduwa University, Ipetumodu, P.M.B. 5533, Ile-Ife, Nigeria

[2]Department of Mathematical Sciences,
Oduduwa University, Ipetumodu, P.M.B. 5533, Ile-Ife, Nigeria

Copyright © 2014 ISSR Journals. This is an open access article distributed under the ***Creative Commons Attribution License***, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT: An air pollution monitoring application system for analyzing and forecasting air pollutant data was developed in order to provide information about the quality of air we breathe. Two industrial environments were used as case studies namely Ife steel plant and Ibadan Asphalt Company. The application was developed using Microsoft visual studio 2012 for the client side and user interface while MYSQL was used for the database. System flowchart was used to design the application modules. Relevant data were collated from the data acquisition systems in Ile-Ife and Ibadan to develop the application. The application when implemented will enable users living and working in the area of study to know the extent at which the air is polluted, forecast the air data and store the collated data in a relational database which will be updated periodically for analysis. This study will attempt to help individuals to know the quality of air they breathe in any particular environment.

Keywords: Pollution, AQI, Air, Model.

1 INTRODUCTION

Air pollution is the introduction of chemicals, particles, biological materials, or other harmful materials into the earth's surface (Wikipedia, 2001). Air pollutants are airborne particles and gasses that occur in concentrations that endanger the health and well being of organisms thus disrupting the orderly function of the environment (Schultz et al 2009). Air pollution thus occurs when the air contains gases, dust, fumes or odour in harmful amounts.

However, informed observers will find these demands surprising so long as the developing countries are not ready to take responsibilities for managing immediate environmental concerns, such as urban air quality, within their own jurisdiction. Consistent with result from cities around the world, air pollution is attributed to the emission from transport, industry, energy and domestic sources, most places are exposed to significant ambient air pollution concentration due to increased reliance on small-scale petrol powered generators for commercial energy supply, uncontrolled open incineration of waste, major thermal power stations and substantial petrochemical activity within the city limits.

Nigeria is faced with so many types of air pollution but the main concern of this project work is the pollution of air caused by industrial emission. All air emission has a hazardous effect on human system such as Ozone and particles matter which causes a wide range of adverse health effects.

Ground-level ozone (O_3) is a product of nitrogen oxides (NOx) and volatile organic compounds (VOCs) in the presence of heat and sunlight. Motor vehicle exhaust, industrial emissions, gasoline vapors, and chemical solvents are among the major sources of NOx and VOCs responsible for harmful build up of ground-level ozone. Even at low concentrations, ozone can

Corresponding Author: E. O. Ofoegbu

trigger a variety of health problems such as lung irritation and inflammation, asthma attacks, wheezing, coughing, and increased susceptibility to respiratory illnesses.

Particles matter (PM), or airborne particles, include dust, dirt, soot, and smoke. Some particles are directly emitted into the air, for example, cars, trucks, buses factories, construction sites, and wood burning. Other particles are formed in the air when gases from burning fuels react with sunlight and water vapor. Such gases, from incomplete combustion in motor vehicles, at power plants and in other industrial processes, contribute indirectly to particulate pollution. PM can cause chronic bronchitis, asthma attacks, decreased lung function, coughing, painful breathing, as well as a variety of serious environmental impacts such as acidification of lakes and streams and nutrient depletion in soils and water bodies.

2 REVIEW OF RELATED LITERATURES

Wong Tze Wai (2009) studied an air pollution index reporting system. The Air Pollution Index (API) Reporting System is an important tool of communication risk. It informs the public of ambient air pollution, and the potential health risk, it imposes on groups such as children, elderly, and those with cardiovascular and respiratory diseases. People use the API to make decisions on outdoor activities; such as, schools and sports organizations, sports organization uses it to check the latest API figures to decide whether outdoor sporting events should be conducted on a certain day. An air pollution index system was developed in Hong Kong, with full justifications and implementation details.

WHO published a Monitoring ambient air quality for health impact assessment. The study shows the importance of the availability of valid information on population exposure to air pollutants, European Centre (WHO) for Environment and Health, organized a working group with the objective of defining the features of monitoring networks that gives room for assessing the potential exposure of the population to air pollution from ambient air. The air quality assessment must include links with population exposure and the pollution sources. The principles outlined in these report are intended to promote progressive modification of the networks monitoring air quality, to improve their usefulness for health impact assessment.

Gupta (2008) studied air quality in an urban region of Kolkata, which consist of residential, commercial and industrial sites, having high population density and pollution. Concentrations of ambient SO_2 (sulfur dioxide), NO_2, (nitrogen dioxide), NH_3 (ammonia) and PM_{10} was measure at selected residential, industrial sites and commercial site. The meteorological parameters (wind speed, wind direction, rainfall, temperature and relative humidity) were collected simultaneously from the Indian Meteorological Department, Kolkata. Winter concentrations of ambient such as SO_2, NO_2, NH_3 and PM_{10} were observed to be higher irrespective of the monitoring sites and duration of sampling.

Defra (2010) studied air quality across UK and EU, the report showed that the air quality index has increased in the improved last couple of decades. However, there are still some evidence of negative health effects and environmental damage caused by emissions of air pollutants such as Particulate matter (PM), Ammonia (NH_3), Oxides of Nitrogen (NOX) and Sulphurdioxide (SO_2). The study showed that the major air pollutants of concern in UK are PM, NOX, ozone (O_3) and NH_3. Meanwhile, climate change has emerged as a major global challenge with achievement of legally binding targets by 2050, a key priority for the UK Government and the devolved administrations (DAs). Across Government, work is now underway through the 2009 Low Carbon Transition Plan (LCTP) 1 to meet the carbon budget commitments from 2008 through to 2022.

Schultz R. B (2006) studied air pollution and weather pollution. Weather is linked in two ways. The first way is concerns about the influence that weather conditions have on the dilution and dispersal of air pollutants. The second way is the reverse which deals with the effect that air pollution has on weather and climate. Air is never perfectly clean, Examples of natural air pollution include: Ash, salt particles, pollen and spores, smoke and windblown dust. The study show the type of air pollution which are primary pollutant and secondary pollutant the major primary pollutants are: Particulate matter (PM), sulfur dioxide, nitrogen oxides, volatile organic compounds (VOCs), carbon monoxide, and lead. The **Clean Air Act of 1970** mandated the setting of standards for four of the primary pollutants Particulates, sulfur dioxide, carbon monoxide, and Nitrogen as well as the secondary pollutant ozone The study show that primary pollutants in the United States are about 31 percent lowers than 1970. In 1990, Congress passed the **Clean Air Act Amendments**, which further tightened controls on air quality. Regulations and standards regarding the provisions of the Clean Air Act Amendments of 1990 are periodically established and revised.

Jong-won kwon (2007) designed and implemented an air monitoring system using sensors. Each sensor was tested after survey for detecting air pollution. Secondly, wireless communication modules for monitoring system were developed using wireless sensor network technologies based on ZigBee and performance of modules was estimated in the real-fields. Through software program written in C, efficient routing in wireless networks was simulated using the TOSSIM simulator. Finally,

integrated wireless sensor board, which employs dust, CO_2, temperature/humility sensor and ZigBee modules was developed. The work accelerates the digital convergence age through continual research and development of technologies.

Kavi (2005) designed and implemented a wireless sensor network Air pollution Monitoring system. The work investigates the use of wireless sensors networks for air pollution monitoring in Mauritius. With the fast growing industrial activities on the island, the problem of air pollution is becoming a major concern. The paper proposed an innovative system named wireless sensor networks air pollution monitoring system to monitor air pollution in Mauritius through the use of wireless sensors deployed in huge numbers around the island. The system makes use of an Air Quality index in order to improve the efficiency of the system.

Casella (2004) designed software called Envista Air Resource Manager Software (ARM), Envista Air Resource Manager software (ARM) is a client-server application for supervisory control, management and analysis of data from Environmental, Meteorological and Hydrological monitoring networks. It is a powerful air quality software suite and has been designed to meet the needs of both air quality experts and general users alike. It's simple, easy-to and allows any user to produce professional tables or graphs. Customized dashboards can also be constructed and personalized data reports can be generated, files can be immediately converted into a CSV format and imported into other statistical packages.

3 MATERIALS AND METHOD

This section presents a detailed explanation of the research methodology adopted in this study. The materials that were consulted are briefed and also the method used to achieve the objectives of the system are given step-wisely.

3.1 DATA COLLECTION AND PROCESSING

This is the next phase after understanding the research problem. Business data are stored using different types of systems across an enterprise. The first step is to pull the relevant data to a database or data mart where the data analysis is applied. The data used in the context of this Research were samples of air pollution data from two polluted environments which are Ife steel plant and Ibadan Asphalt Company.

The data used for this project were grouped in excel file format based on the pollutant and the city being considered, thereafter the data in an excel format was uploaded on the SQL (Structured Query Language) Server for analysis. This operation was made easy as a result of the flexibility of the .net environment (Visual Studio) using LINQ (Language Integrated Query) to Excel.

3.1.1 DATA CLEANING AND TRANSFORMATION

Data cleaning and transformation is the most resource-intensive step in a data analysis project. The purpose of data cleaning is to remove noise and irrelevant information out of the dataset. The purpose of data transformation is to modify the source data into different formats in terms of data types and values. There are various techniques that can be applied to perform data cleaning and data transformation, *Data Type* Transform: this is the simplest data transform. An example is transforming a Boolean column type to integer. The reason for this transform is that some data analysis algorithms perform better on integer data, while others prefer Boolean data.

Much of the raw data contained in excel sheet were un-pre-processed, incomplete, and noisy. For example, the data in excel format initially contained fields that are obsolete or redundant, missing values, outliers, data in a form not suitable for input to our analysis model, and values not consistent with policy or common sense.

Table 1: Air Pollution Data

Year	So_2			No_2			Pm_{10}		
	Year	Month	Day	Year	Month	Day	Year	Month	Day
2000	80	125	1500	50	220	1000	100	25	
2001	80	125	350	45	85	846	125	21	5
2002	40	846	500	100	876	400	150	10	
2003	67	365	260	54	76	590	80	150	
2004	32	365	780	23	87	760	60	120	
2005	76	400	2620	43	150	320	150	30	30
2006	86	65	1500	51	50	65	200	20	
2007	43	150	550	55	40	55	175	10	
2008	90	125	743	49	120	300	140	5	
2009	20	50	1500	62	360	570	150	20	
2010	15	30	350	80	67	70	90	150	10
2011	54	87	540	85	120	80	120	10	
2012	50	341	1350	78	56	301	150	30	
2013	43	62	120	55	110	45	75	140	20
2014	61	120	110	70	125	400	50	120	

3.1.2 HANDLING MISSING DATA

Missing data is a problem that continues to plague data analysis methods. Even as our analysis methods gain sophistication, missing values in fields, especially in databases with a large number of fields are often encountered. The absence of information is rarely beneficial. All things being equal, more data is almost always better. Therefore, in order to handle the thorny issue of missing data, the missing value is replaced with:

a. some constant, which could be specified by the analyst;
b. the field mean (for numerical variables) or the mode (for categorical variables); or
c. a value generated at random from the variable distribution observed.

3.2 DATA FORECASTING MODEL

A data forecasting model or forecasting model can be thought of as a relational table. It contains key columns, input columns and predictable columns. Each model is associated with a forecasting algorithm on which the model is trained. Training a forecasting model means finding the pattern in the training dataset by using specified data forecasting algorithms with proper algorithm parameters. After training the model we store the patterns that the data forecasting algorithm discovered about the dataset. While a relational table is a container of records, a data forecasting model is a container of patterns. The system predicts three air pollution indicator levels for the next one hour, twenty-four hour or even a year. The inputs were general condition, wind direction, pressure, day temperature, night temperature, relative humidity and wind speed. Output parameters were sulfur dioxide, particulate matter and carbon monoxide.

3.2.1 MODEL CREATION

The concept of model creation simply deals with creating an empty data forecasting model, similar to the way a relational table is created.

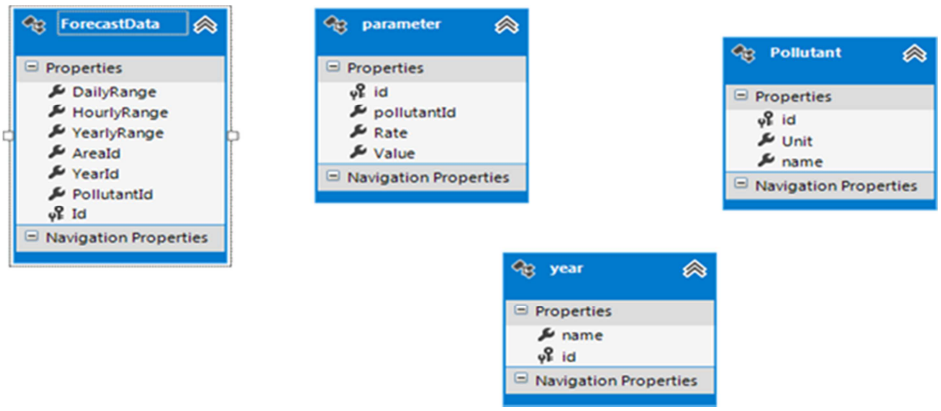

Figure 1: Diagram of the Relational Database Model Adopted

3.2.2 MODEL TRAINING

Model training is also called model processing. It is used to invoke the data forecasting algorithm to uncover knowledge about the training dataset. After training, patterns are stored in the model. Artificial neural network is used to train the network and forecast pollutant data for consecutive years. The network consists of 3 inputs, 4 hidden and 2 outputs. The input to the network model are pollutant data from previous years while the hidden layer consists of a sigmoid function and an hyper tan function that ensures the weight and biases helps to generate a reasonable pollutants data. Backward propagation medium is adopted to train the network and reverses its learning process if the needs arise. At the output end two sections of pollutant data are produced and the decision tree algorithm is then applied to determine the most viable data to be used as a forecast.

3.2.3 DECISION TREE ALGORITHM

Decision trees are powerful and popular tools for classification and prediction, decision trees represent rules, which can be understood by humans and used in knowledge system such as database. The principle idea of decision tree is to split the data recursively into subsets so that each subset contains more or less homogenous states of your target variable (predictable attributes). At each split in the tree, all input attributes are evaluated for their impact on the predictable attribute.

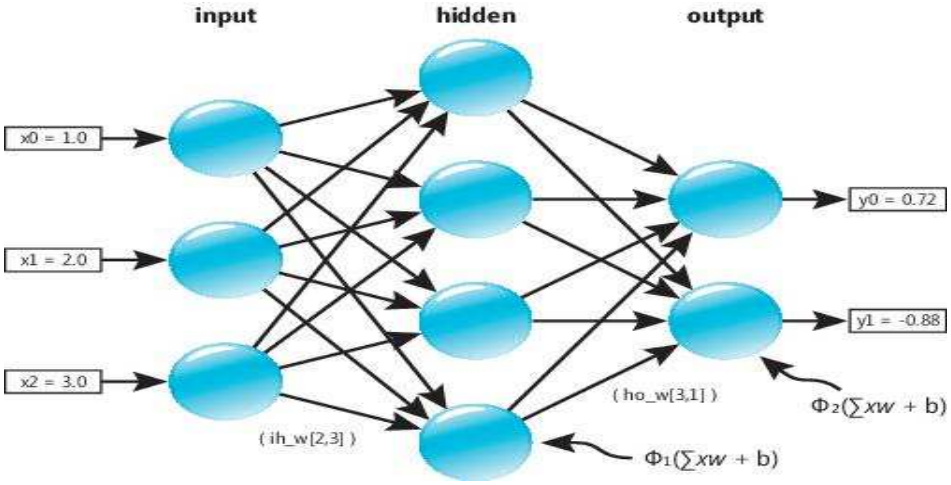

Figure 2: ANN Model

When this recursive process is completed, a decision tree is formed. There are few advantages of using decision trees over using other data forecasting algorithms, for example, decision trees are quick to build and easy to interpret. Each path from the root forms a rule. The key requirements of decision tree are:

a. **Attribute-value description:** Object or case must be expressible in terms of a fixed collection of properties or attributes (air quality i.e. good or bad).
b. **Predefined classes (target values):** the target function has discrete output values (Boolean or multiclass)
c. **Sufficient data:** enough training cases should be provided to learn the model.

Decision tree learning uses a decision tree as predictive models which maps observation about an item to conclude about an items target value. It is one of the predictive modeling approaches used in statistics, data mining and machine learning, in decision analysis tree can be used to visually and explicitly represent decisions and decision making. In data mining, decision tree describe data but not decisions; rather the resulting classification tree can be input for decision making. Decision tree is a method commonly used in data mining; the goal is to create a model that predicts the value of a target variable based on several input variables.

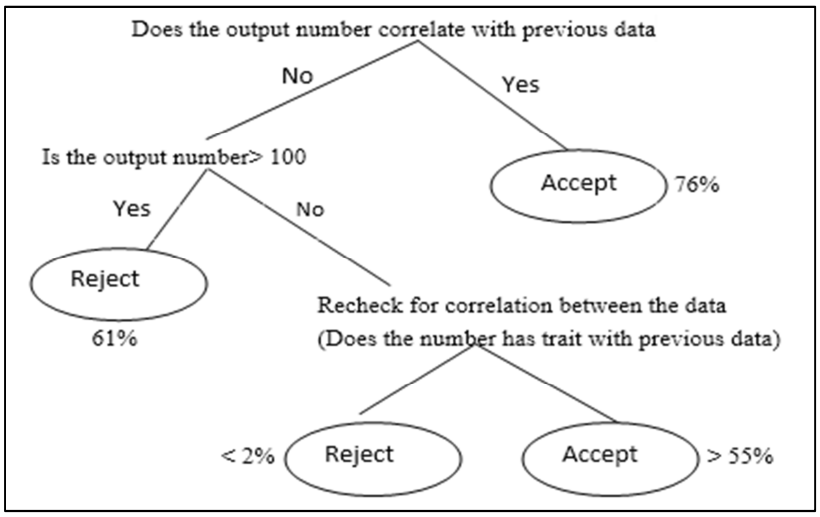

Figure 3: Decision Tree Used to Select Between the Two Output Values

3.3 SYSTEM ANALYSIS AND DESIGN

The air pollution data forecaster was written in a more interactive form, so as to help a user relate with the application more effectively. The front-tier of the application program developed was compiled in Microsoft Visual Studio 2012 framework while the SQL Server 2008 was used for the back end. This is to keep the data before the stage of pre-processing and analysis. The front tier has different modules which are linked together.

SYSTEM FLOW CHART

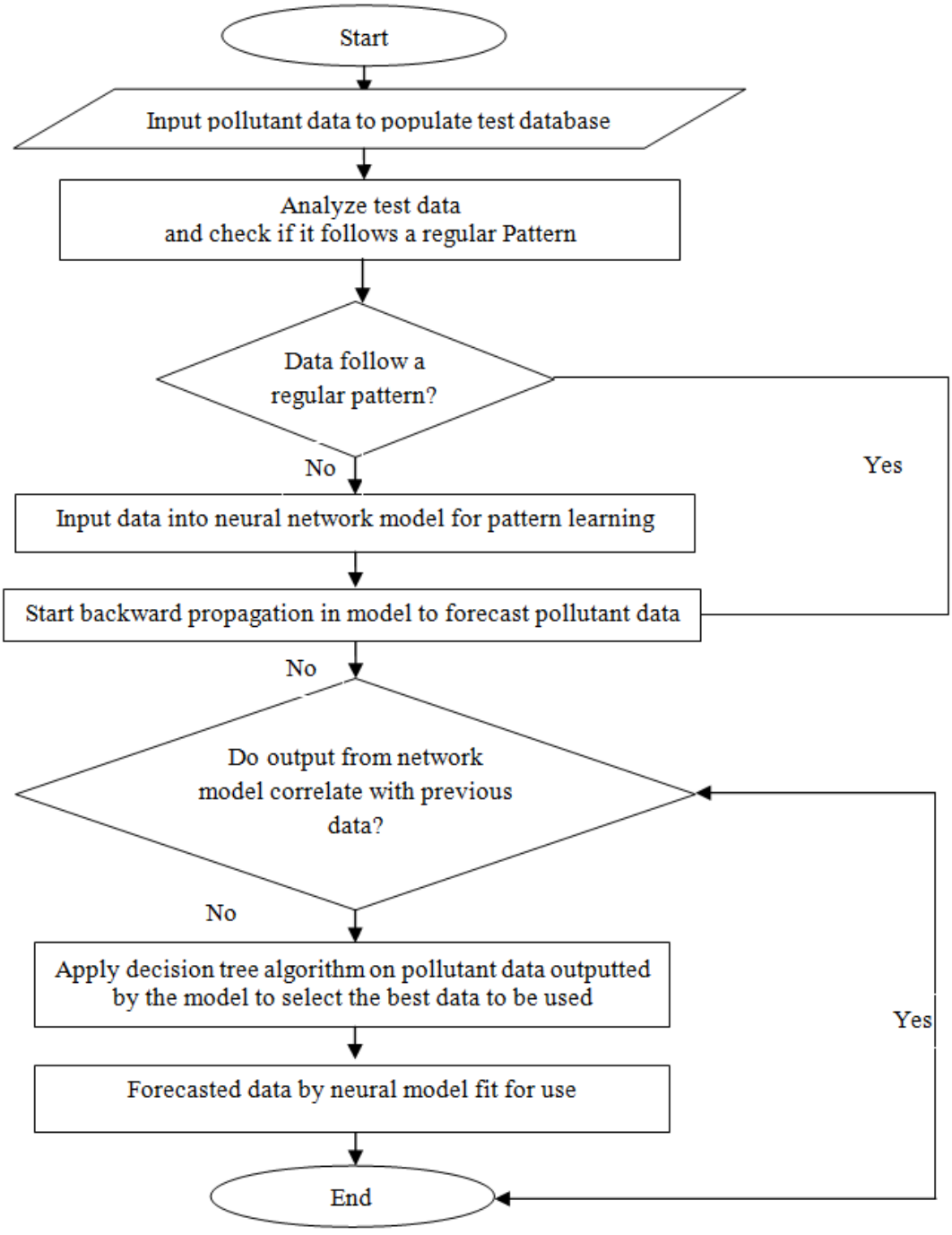

Figure 3: System Flow Chart

4 SYSTEM IMPLEMENTATION AND DISCUSSION

Implementation was done using Microsoft visual studio for the server controller (linkage between client and server) and data's were extracted from the database using SQL. The diagrams below shows interface of the system.

a. Add Pollutant Details: This section enables the user to add pollutant details which would enable the system to determine the kind of pollutant data to analyze. Pollutant name could be So_2,No_2 etc while its unit could be ug/m3 and mg/m3.
b. Add area details: This section enables the user to add area/s that would be considered during our data analysis.

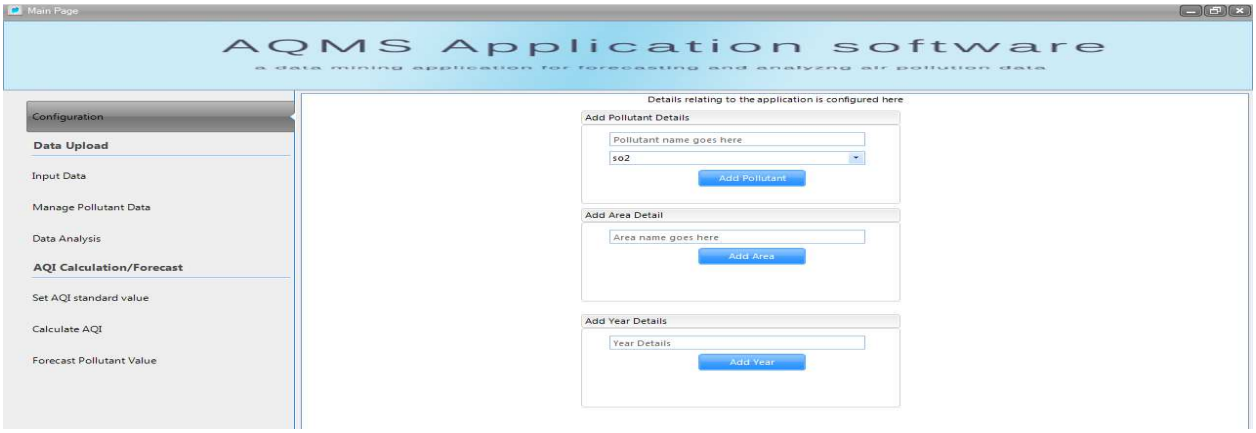

Figure 4: Configuration Section

c. Add year details: This section enables the user to add years that would ensure pollutant data for year added would be supplied.

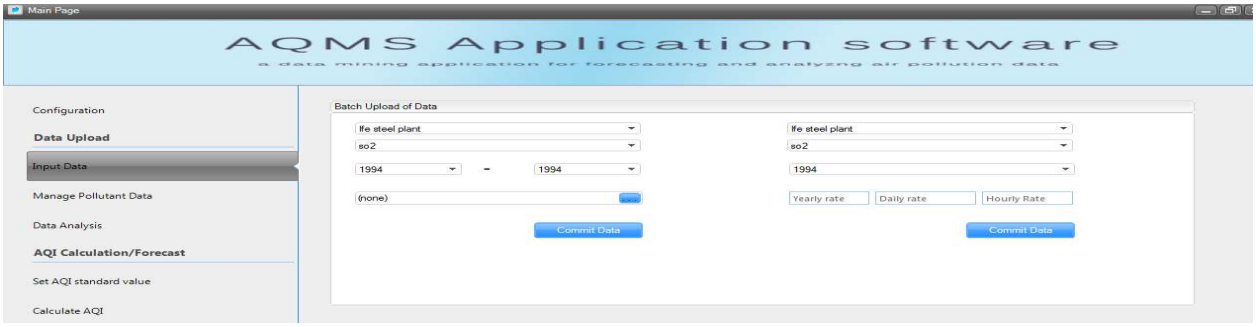

Figure 5: Batch Upload of Data

Data upload section enables the user to supply pollutant data that would ensure the system to perform its analysis and forecasting functionality, Data to be upload can be in form of an excel file or supplied individually under a selected year, area and particular pollutant

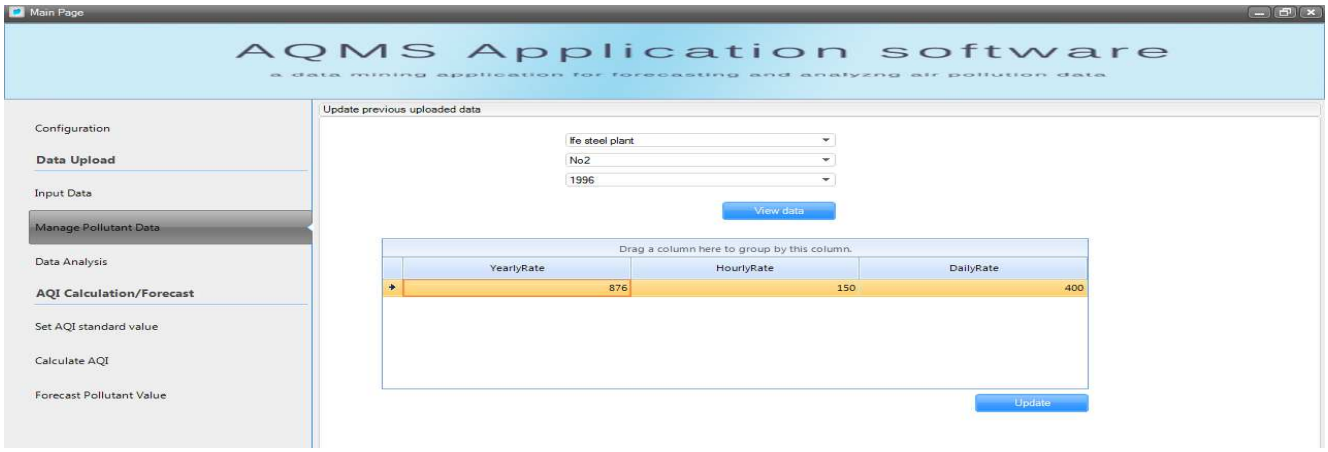

Figure 6: View Uploaded Data

This section allows the user to view previously enrolled data, to check for consistency in them and also change data values if there is need to.

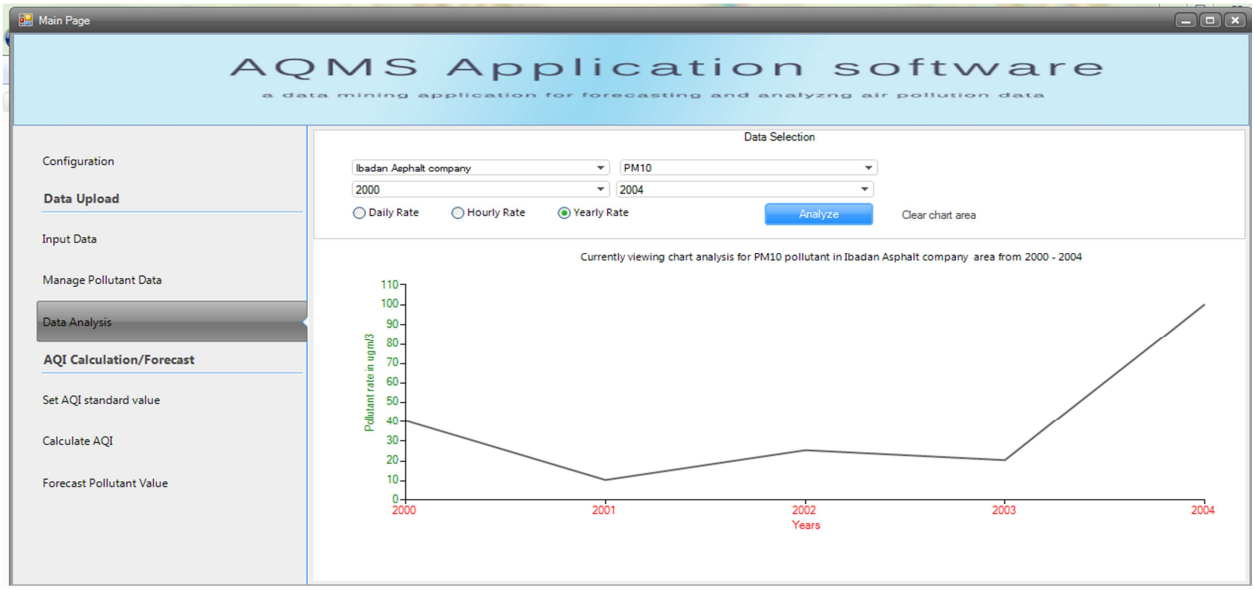

Figure 7: Data Selection and Analysis

This section provides a graphical interface that helps to analyze pollutant data from range of years selected and also area specified

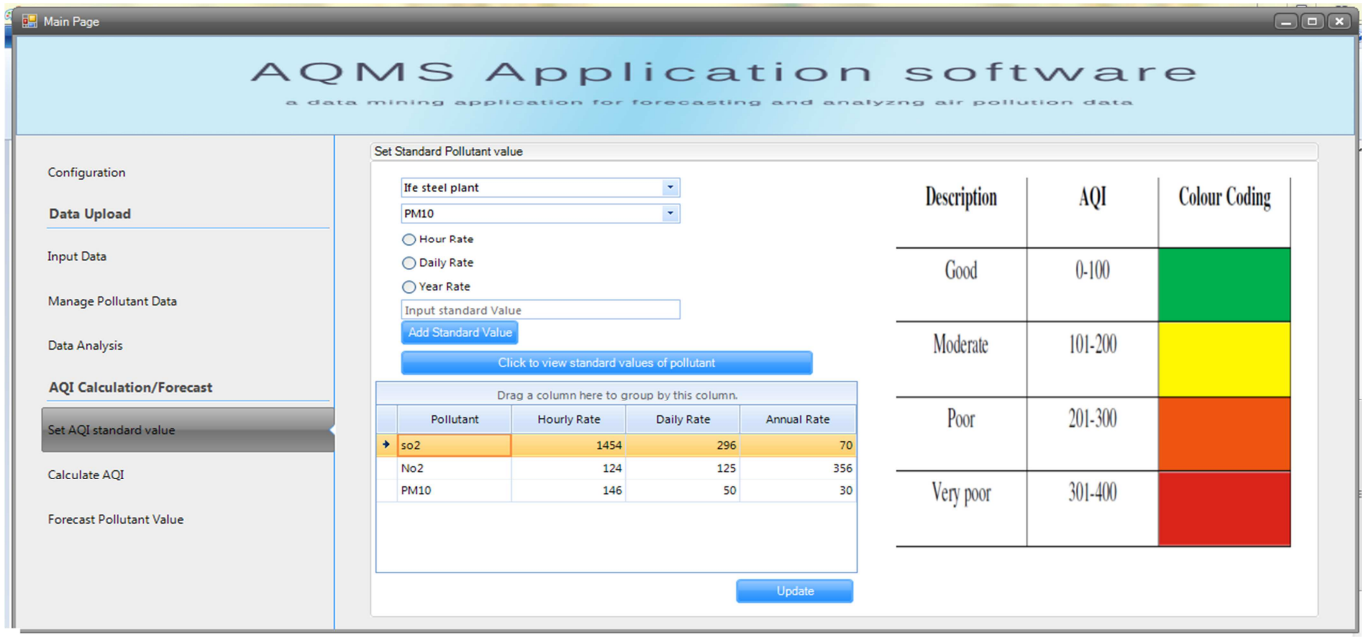

Figure 8: Set Standard AQI Pollutant Value

This section allows the user to supply standard value for pollutant data across diverse areas, hence ensuring a correct AQI value is calculated. It also provides a means of updating previous standard values.

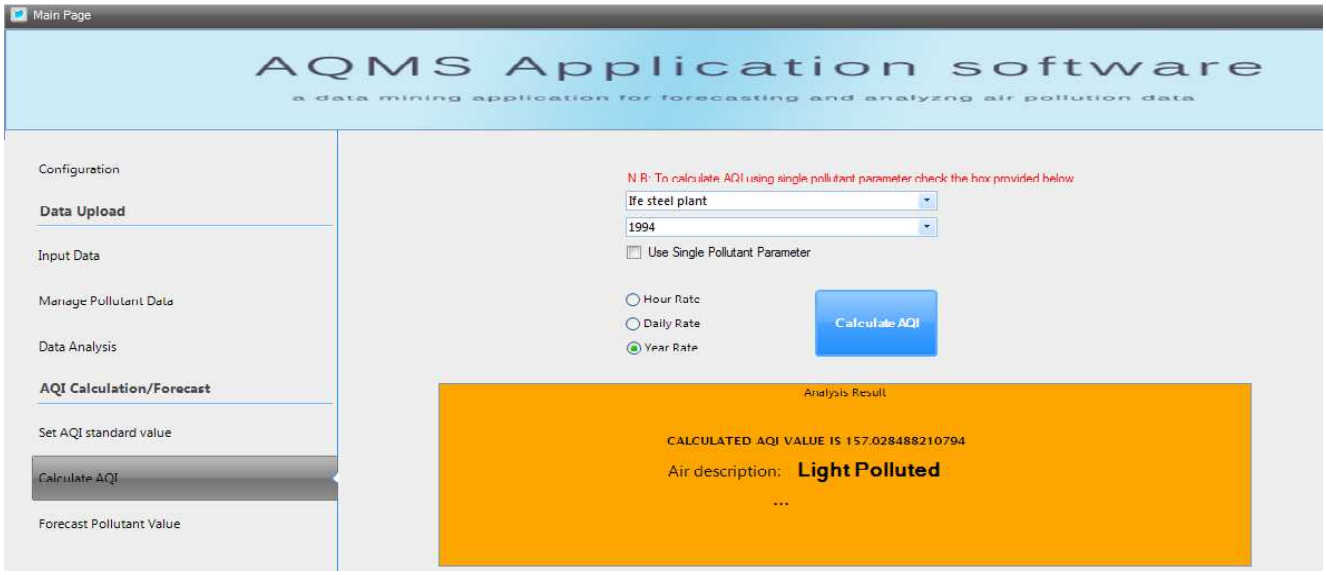

Figure 9: Calculation of AQI

This section provides a means to calculate AQI value of pollutant rate in a particular area in a specified year. Detailed description of the result is also displayed for user consumption.

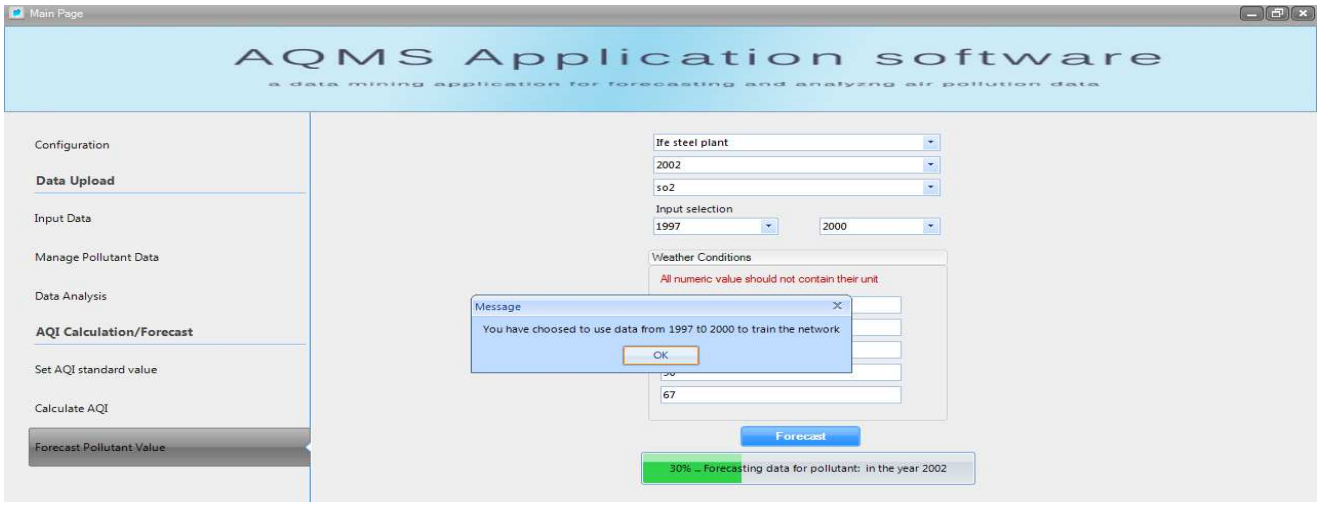

Figure 10: Forecast Pollutant Value

This section allows the system to perform its basic functionality which is to forecast pollutant data for a particular area in a specified year. The viability of the forecast result is a function of the input data from previous years

When forecasting weather conditions are put into consideration, because it affect the rate of pollution, all weather conditions are inputted, and the system forecast data for the selected year.

5 CONCLUSION AND FUTURE WORKS

Air pollution is the introduction of chemicals, particles, biological materials, or other harmful materials into the earth's surface (Wikipedia, 2001). Nigeria is faced with so many types of air pollution, but the main concern of this research effort is air pollution caused by industrial emissions. All air emission has a hazardous effect on human system such as Ozone and particulates matter, which causes a wide range of adverse health effects. The **AQMS** application software performs the analysis of pollutant information to calculate air quality index and for forecasting data, it allows people in a particular

location to monitor the quality of air they breathe in. Using this software, pollution rate can be reduced and people with cardiovascular disease will know the environment conducive for them to live.

In the light of increase in air pollution, it is apparent that there is no functioning air monitoring system in the country. It is recommended that building new systems especially for Air quality measurement is very essential and crucial.

This paper can also enable the development of other air monitoring system in other states to be developed which can be done via any programming language when data are generated. I recommend that further works should be done especially with the help of professionals in engineering and computer science while considering other pollutant in future work. Therefore, it is suggested that applications that run on web enabled mobile devices is incorporated to access air quality in different locations.

REFERENCES

[1] Air Pollution Definition Retrieved from Wikipedia (07-07-2014)

[2] Boehm Spiral Model Software Development process retrieved 08-06-2014

[3] Chattopadhyay S., Gupta S. and Saha R. N. (2010), "Spatial and Temporal Variation of Urban Air Quality A GIS Approach". Journal of Environmental Protection, 1: 264-277.

[4] Chelani A. B., Chalapati Rao C. V., Phadke K. M. and Hasan M. Z. (2002), "Formation of an Air Quality Index for India". International Journal of Environmental Studies, 59: 331-342

[5] Casella (2004)" Air quality monitoring: it's all connected, a solution based approach". www.Casellameasurement.com

[6] Clean air Act 1970 retrieved from Wikipedia (2013) "how air quality data is used".

[7] R. B Schultz (2010). "Air pollution"

[8] Defra (2010) "Air Pollution Action in a Changing Climate" retrieved (09-07-2014)

[9] Gupta A. K., Karar K., Ayoob S. and John K. (2008), Spatio-Temporal Characteristics of Gaseous and Particulate Pollutants in an Urban Region of Kolkata, India. Atmospheric Research, 87: 103-115.

[10] Gufran B., Ghude D. S. and Deshpande A. (2010), "Scientific Evaluation of Air Quality Standards and Defining Air Quality Index for India" Indian Institute of Tropical Meteorology Research Report No. Rr-127

[11] Kavi K. Khedo, Rajiv Perseedoss and Avinash Mungur (2005). "A Wireless Sensor Network Air Pollution Monitoring System" Department of Computer Science and Engineering, University of Mauritius

[12] Bharati M. Ramageri "Data Mining Techniques and Applications" Indian Journal of Computer Science and Engineering 301-305

[13] Wong Tze Wai : "A Study of the Air Pollution Index Reporting System" School of Public Health and Primary Care, The Chinese University of Hong Kong

[14] Upadhyaya G. and Dashore N. (2010), "Monitoring of Air Pollution by Using Fuzzy Logic". International Journal on Computer Science and Engineering, 2: 2282-2286

[15] WHO (2000), "Monitoring ambient air quality for health impact assessment" (WHO regional publications. European series; No. 85)

[16] Whitten Jeffrey (2003) ……"what are RAD model-advantages, disadvantages and when to use it?" ISTQBEXAM CERTIFICATION RETRIEVED (08-06-2014)