Mining of Air Pollution and Visualization

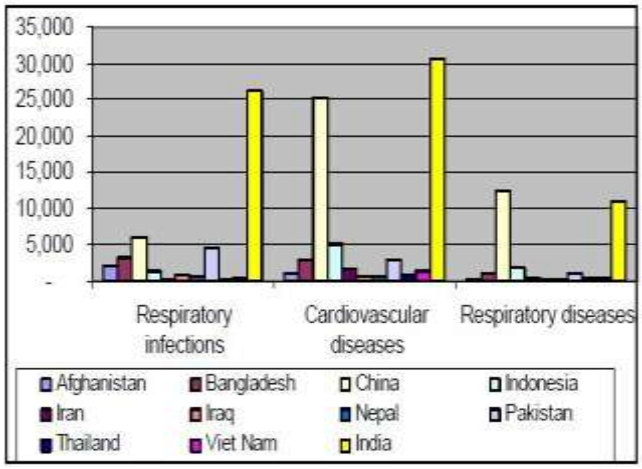
Kriti Gupta
Department of Information Technology
Poornima Institute of Engineering and Technology
Jaipur, Rajasthan, India

Abstract—In this growing era of 'Smart Cities', there are certain factors which are being neglected. One of them is Air Pollution that is causing so many rigorous diseases in humans. The major pollutants are SO2, NO2, and PM10. So we will be finding the cities where air pollution due to these pollutants has already crossed its dangerous mark and the cities having average and low level of air pollution through Data Mining. Data Mining is a process through which certain useful and interesting patterns are extracted from the huge dataset and on the basis of them, some decisions can be taken. Patterns extracted from the mining procedure will be used to visualize the pollution and may help our government and society to take much crucial steps towards the highest polluted areas.

Keywords—*data-mining, patterns, smart cities, rigorous diseases, dangerous mark.*

I. INTRODUCTION

In developed as well as developing countries, the increased levels of air pollution are a major environmental problem. Pollution has become a great topic of concern at all levels in India and especially the air pollution because of the enhanced human and anthrop genetic activities. Among the harmful chemical compounds entering into the atmosphere as a result of burning of fossil fuels, are Carbon monoxide (CO), Sulphur Dioxide (SO2) Nitrogen Dioxide (NO2), and tiny solid particles–Particulate Matter including lead from gaseous additives. The studies on air pollution in large cities of India showed that terrain air pollution concentrations are at such levels where serious health effects are possible. Continuous rise of population due to urban activities along with the lack of suitable measures for controlling the air pollution means that there is a great potential that conditions may get more worse in future in Indian cities. In the urban area the air quality is affected adversely due to emission and agglomeration of PM10, SO2, CO and NO2 [1]. These all pollutants may cause harmful effects on human health, as exposure of these are associated with cardiovascular and respiratory disease, Neurological impairments, increased risk of premature birth and even mortality and vexation. Various studies conducted in India at various locations suggest that pollution levels varies significantly in different areas with reference to its location, time and climatic conditions.



Source: World Health Organization, Department of Measurement and Health Information, December, 2004

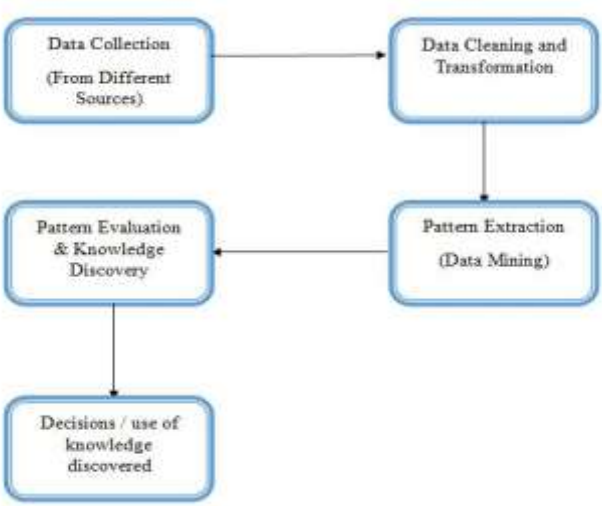
This figure shows the amount of deaths that are caused due to air pollution across various countries. As we can see, in India the amount is much higher than other countries. It has been found that SO2 and NO2 concentrations are within the permissible limits in many areas but SPM and PM10 concentrations are generally exceeding the limits as per the guidelines given by Indian Air Quality (National Ambient Air Quality Standard-2014)[2].

Among the pollutants listed in NAAQS, one of the most prominent pollutants is PM10. It is well known that PM10 is responsible for respiratory hazards in human health. Such particulates can also restrain lung functions without reacting chemically, by depositing in human lungs and interrupting with normal functioning. More often, it takes part in formation of sulphurous haze. One of the main sources of existence of PM10 in air is pollution caused by vehicles [3]. Various typical phylogenies activities like intense transportation, Industrial and commercial activities are predominating in urban areas, particularly in the cosmopolitan cities. It is also known that increased level of primary pollutants like Sulphur-dioxide (SO2) and Nitrogen-dioxide (NO2) lead to the formation of different types of secondary pollutants in environment. Studies reveal that the occurrence is mainly due to

expansion of industries and growing number of vehicles within the country.

Cities have always been considered engines of industrial growth, as they offer their residents opportunities for employment and prosperity. This fact has become particularly pronounced in modern times. There is an extreme need of taking more such steps and for that we must have the measures and parameters of air pollution.

II. Architecture And Methodology



III. Data Collection

This is the next step to be performed after understanding the problem. The data is taken from the government site year-wise and city-wise.

City 2013	SO_2	NO_2	PM_{10}
Agra	5	21	184*
Ahmadabad	12	17	79*
Allahabad	5	29	235*
Amritsar	13	40	180*
Aurangabad	10	37	84*
Bangalore (BBMP)	13	26	113*
Bhopal	3	26	220*
Chennai	14	22	75*
Coimbatore	4	24	56
Delhi (DMC)	4	66*	221*
Dhanbad	16	40	151*
Faridabad	12	26	196*
Ghaziabad	26	34	285*
Gwalior	13	27	197*
Howrah	11	45*	187*
Hyderabad (GH)	5	24	90*
Indore	11	19	156*
Jabalpur	2	23	69*
Jaipur	7	40	160*
Jodhpur		23	176*

(Source:-cpcbenvis.nic.in/air_quality_data.html)

As we can see, the dataset consists of three parameters SO2, NO2 and PM10. The amount of these pollutants is mentioned in per meter cube.

Similarly, we are having dataset of years from 2016 to 2013.On this dataset, we are going to perform mining and on the basis of this, we will be forecasting some decisions.

IV. Data Preprocessing

Data preprocessing comprises of Data Cleaning and Data Transformation. On observing the dataset, we see that the dataset is having certain missing values and noisy data too. So for further processing, we need to clean that data. This step is known as Data Cleaning.

Dataset is having data in different-different formats. So, we need to reformat the data into some specific format. This step is termed as Data Transformation.

SNO	City	SO2	NO2	PM10	Year
1	Agra	5	21	184	2013
2	Ahmadabad	12	17	79	2013
3	Allahabad	5	29	235	2013
4	Amritsar	13	40	180	2013
5	Aurangabad	10	37	84	2013
6	Bangalore	13	26	113	2013
7	Bhopal	3	26	220	2013
8	Chennai	14	22	75	2013
9	Coimbatore	4	24	56	2013
10	Delhi	4	66	221	2013
11	Dhanbad	16	40	151	2013
12	Faridabad	12	26	196	2013
13	Ghaziabad	26	34	285	2013
14	Gwalior	13	27	197	2013
15	Howrah	11	45	187	2013
16	Hyderabad	5	24	90	2013
17	Indore	11	19	156	2013
18	Jabalpur	2	23	69	2013
19	Jaipur	7	40	160	2013

This is the cleaned data having no missing values, no special symbols. On this cleaned dataset, we will perform the further mining algorithms.

The cleaning is done through R-Programming. In R, there are some predefined functions of removing these unwanted values:

gsub( ):-

```
Cleaded_data<=gsub(„[a-zA-Z0-9.]“ ,data);
```

This function will allow only the values that are passed in it and removes all other values. The second parameter is the data which is to be cleaned.[5]

FindReplace( ):-

It is a function to replace multiple characters found in the data passed in it.

rmExcept( ):-

It removes all the objects from a dataset except those which are explicitly specified by the user.

A user-defined function can be also made for cleaning the whole dataset at once:

```
clean<= function(data)
{
    as.numeric(gsub(„[a-zA-Z0-9.]“ ,data));
}
data[]<= sapply(data,clean);
```

V. PATTERN EXTRACTION

Data mining is a technique used for extracting some interesting and useful patterns from the cleaned dataset that is given by the previous step. Several algorithms are available for performing Pattern Extraction.

In this dataset, the patterns that we have identified are:

Finding the cities which are most and least vulnerable to pollution.

Finding the variations in the amount of pollutants in different cities from last seven years.

Then, predicting the cities where pollution will cross its dangerous level after two years, five years and seven years.

Due to this, we can communicate to higher authorities about these cities so that they can focus more on these cities.

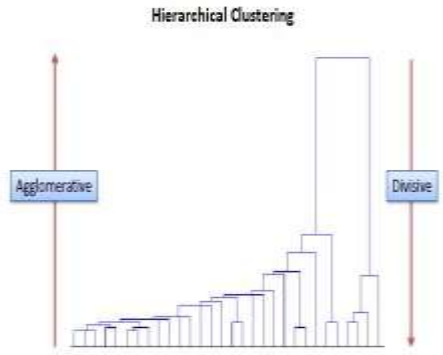
For implementing pattern extraction, first we need to cluster our dataset into groups having proper formats.

VI. DATA CLUSTERING

Data clustering is a process of partitioning the datasets into a set of meaningful sub-classes called clusters. Many algorithms are there for performing clustering process like k-Means clustering algorithm, Hierarchical clustering algorithm, Density based clustering algorithm.

Hierarchical Clustering:

Hierarchical clustering includes creation of different clusters having a predetermined ordering from top to bottom. There are two types of hierarchical clustering: *Divisive* and *Agglomerative.*



K-Means Clustering:

K-Means clustering intends to make partitions of *n* objects into *k* clusters, in which each object belongs to the cluster with the nearest mean. The *k* different clusters produced are having greatest possible distinction.[6]

Steps in K-Means Clustering:

We have performed clustering by using R-Programming:

```
data=read.csv(file.choose()) x<-
rbind(matrix(data,ncol=2),matrix(data,ncol=2))
colnames(x)<-c("x","y")
out=head(model.matrix(~.+0, data))
kmeans(out,4)
```

In step one, a csv file is read from the system and stored in a variable „data“ .

Since, our data is not in the format that is required by R. So, we bind the data with the function „rbind()“ .

Now, we are aliasing the data with the names „x“ and „y“ .

More refining is performed on the dataset by the function „modelmatrix()“ in next step.

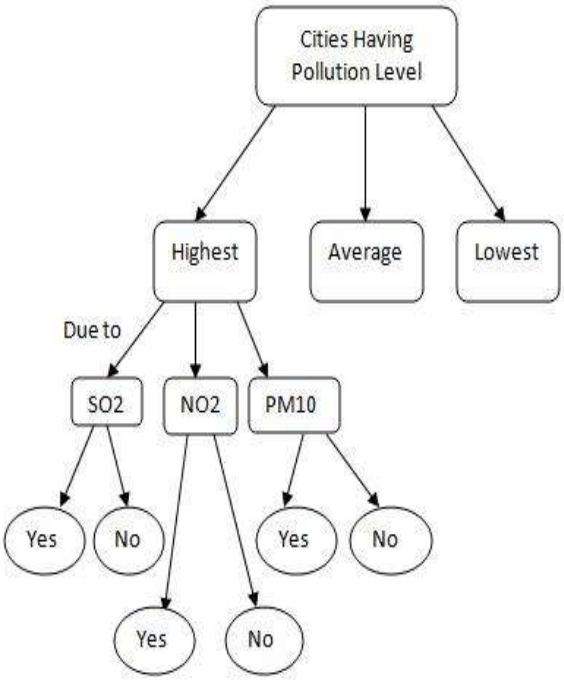
Lastly, function „kmeans()“ is performing clustering on data stored in „out“ variable and 4 represent the number of clusters.

Now, we are having four major clusters of our dataset. On these clusters, we can perform algorithms like Decision Tree Algorithm; Back-Propagation Algorithm.

VII. DECISION TREE ALGORITHM

Decision trees are the prediction and classification tools. The output of decision trees are the rules that are easily understand by the humans and used in large database management systems. The principal concept of decision tree is to split the data recursively into subsets so that each subset contains more or less homogenous states of the targeted variable.

Decision Tree learning is the most widely used and practical methods for inductive inference over supervised data. The formation of decision tree does not require any parameter setting or any domain knowledge and thus efficient for proper knowledge discovery.



VIII. EXPERIMENTAL-SETUP AND RESULTS

To accomplish the task, we have used the large dataset and applied some techniques like data-mining, in which K-Means Clustering is performed to have numerous clusters and decision tree algorithm to find the interesting patterns. Now, for results, we need to visualize the patterns that are extracted from the process stated earlier.

For this, D3 (Data Driven Documents) a JavaScript library is used.

With D3, our extracted patterns can be visualized through different graphs.

In this map, the cities having pollutants concentration are marked. It is a goggle map and by using the functions of D3js,org, the cities are marked on the basis of their latitudes and longitudes. The markers are created by using SVGs.

IX. CONCLUSION

Hence, with the technologies like data-mining and D3, this paper tells us the cities where pollution has already crossed its dangerous mark and the cities where pollution will cross this mark in upcoming years. The cities having highest and lowest amount of SO2, NO2 and PM10 are extracted and thus, the exact reasons for the same can be figured out easily. Due to this, a prediction analysis can be done for the impact of pollution in future „Smart Cities‟ . The result of this prediction analysis can be given to our government so that they can focus much more on these cities in terms of controlling air pollution.

X. 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)