1.1 INTRODUCTION

Understanding and finding out the health status of a society/population is an extremely complex concept, even with supportive epidemiological studies. One of the main reasons for this complexity lays in the multi-factorial nature of health related exposure vectors. Some of these factors are nutrition, psychological and social conditions and the overall indoor and outdoor environmental conditions of the population under study. Extensive literature exists on the effects of these factors on the health of the populace based on which the exposure response functions have been evaluated.

The above argument implies that promoting health of people or improving their quality of life goes hand-in-hand with their interaction with immediate environment. In turn, improvement of environment (indoor and outdoor) leads to prevention of environment-related diseases and thus good health of a populace. In the long run, good health of a population has a direct effect on the economy and educational status of a society. This brings out the importance of a clean environment with investments in public health and science education as part of good governance.

Air Quality

The surrounding -- ambient -- air, generally termed as atmosphere, has been taken for granted by humans. Breathing in and out, billions of time from birth till our death, the quality of air has a direct impact on human health, and also on other life forms on this planet. The air quality keeps on changing on a daily and even on an hourly basis. The development process and increasing industrialization has led to considerable damage and pollution of the atmospheric layer. To improve air quality, it is important that we have a scientific understanding about the atmosphere and take steps to prevent its further deterioration.

Earth’s atmosphere is composed of many layers, based on the changes in the vertical temperatures as follows:

**Troposphere** - It is the lowest layer (up to 10 km) of the atmosphere nearest to earth. It is the storehouse of all major air pollutants, in particular the mixing layer, 1-2 kms height from the earth’s surface.

**Stratosphere** - It extends. From the troposphere to about 50 km. being less dense than the troposphere, there is less mixing of molecules here. Pollutants in this layer reside for
extended periods, leading to long-term global hazards. Stratosphere also contains the ozone layer, which absorbs the UV radiation, thus protecting life on the earth. Exposure to UV is known to cause skin cancer in humans.

**Mesosphere** - It occurs above the stratosphere extending to a height of 80 km. Low temperatures (often as low as \(0 \text{°C}\)) are observed here.

**Thermosphere** - It occurs above the mesosphere. Here temperature increases with height. The atomic oxygen present in the upper part of this layer filters ultraviolet (UV) radiation. This is important as the harmful UV rays could otherwise enter into the inner layers of the atmosphere.

The atmosphere comprises of a mixture of, gases held to the earth by gravity. A sample of dry air from any region at ground level shows the following major constituent gases: about 79% nitrogen, 20% oxygen and 1% of other gases, including water vapor and carbon dioxide.

Besides these gases, the natural atmosphere also contains gases (water vapor, carbon dioxide and ozone) whose concentrations keep on varying due to various processes, both natural and due to human activities (anthropogenic). In addition, non-gaseous constituents like smoke, dust and sea salt particles are also present.

Atmosphere is a finite, non-renewable resource that has to be conserved and treated with care. Unfortunately, today it is being used as a dumping place for several gases, endangering our health and existence on this earth.

**What is Air Pollution?**

Air pollution is the presence of foreign substances in air. Contaminants, which interact with the environment to cause toxicity, disease, physiological defects and environmental degradation, are labeled as pollutants. In fact, some scientists consider any phenomenon (even fog) or substance causing inconvenience to humans as a pollutant. The various sources of pollutants can be broadly classified as:

**Natural sources:** forest fires, sea spray, volcanic eruptions, and from the earth crust. Here the pollutants are: fog, mist, pollens and bacteria.
**Anthropogenic sources:** emissions from industry, automobiles, power stations and smelters, waste incineration, biomass burning, etc.

Combustion sources: stationary (industry) or mobile (vehicle).

- Here the pollutants are: aerosols (particulates), dust, smoke, and fumes and gases and vapors of CO, SO$_2$, NO$_2$ and volatile organic compounds (VOC).

**Water Quality**

All aspects of water, its quality, quantity, its uses and sources, are at the center stage of discussions (academic, social and political) today in the country. On one hand, water quality is steadily deteriorating, and, on the other hand, there are increasing demands on this natural resource. Balanced development demands a scientific understanding of this valuable resource, as water quality and quantity has a direct impact on our health.

Water is an important and the most abundant resource in the biosphere. The water molecule is represented as H$_2$O, with an oxygen atom in the central position, bound by two hydrogen atoms.

The biological importance of water in supporting life is due to its unique physical and chemical properties. Water has maximum density as a liquid at $4 ^\circ$C with highest boiling and melting points, these properties are attributed to the structure of the water molecule. To give one instance, the last property allows organisms to live below the ice cover in denser and warmer waters.

The world’s water can be classified as fresh and saline water. It exists in three different states: liquid (salt and fresh), solid (fresh) and vapour (fresh). Fresh water constitutes only about 3% of the total water, and a large part of this is locked up as ice caps and glaciers. We can find other fresh water in ponds, lakes, rivers and as sub-surface ground water aquifers. This freshwater balance is greatly dependent on the monsoon system in each region. The marine (salt) environment constitutes the large oceans and seas. The salt waters are equally important as they support an entire marine water ecosystem. Pollution of water sources/bodies could be hazardous for our very survival. This finite resource, which is so vital to all the life on this earth, must be protected, conserved and treated with care.

**What is Water Pollution?**
Rapid urbanization, industrialization and certain agricultural practices have led to pollution and deterioration of natural water bodies. These water bodies are constantly being abused by activities, such as, washing and domestic / industrial waste loads. Indiscriminate use of pesticides and fertilizers, combined with inadequate training of farmers and workers, has led to highly contaminated agricultural runoffs being released in water bodies. The highly toxic chemicals (malathion, lindane, DDT and chlorpyrophos) used as pesticides in India, pollute our water bodies, or find their way in the ground water. Several of these chemicals, especially DDT, are banned in most of the countries in the world.

In urban areas, the situation is worsened often by the direct discharge into the natural water bodies of untreated or partially treated sewage and industrial wastes. Water pollution can also result from natural processes, such as, surface runoffs due to rains or presence of dead organic matter, but these are slow processes, giving enough time for the water body to rejuvenate.

Drinking water, which has been treated by municipalities, often is contaminated with sewage (and other waste) in transit from the storage reservoirs to residential units.

Some of the noticeable signs of water pollution are: dark, dirty color of the water body, reduction in transparency, unpleasant smells, unchecked growth of weeds, decrease in the number of fish, oil and grease floating on the surface of the water body and bad taste of drinking water.

Under this study, an initiative has been taken to study the quality of environment by monitoring the air and water parameters in specific sites in Navi Mumbai area. The study aims to assess the pollution-related health impairment / risks among residents in and around the study sites through personal interviews and survey.

These health problems are mainly in the realm of respiratory, gastro-intestinal. These vector borne diseases could partly be attributed to the unhygienic environment with stagnant water pools and open dumping of garbage in public localities that promotes the proliferation of mosquitoes, rodents and stray dogs.

To discover the new knowledge, techniques of data mining could be efficiently applied on spatial data. Classification and regression algorithms are used to understand the data behavior patterns. Predictive analysis is used to evaluate the future trend of data and GIS is used to study its impact on human beings which will be useful for government personnel to take remedial actions towards
the environment, indirectly health victims can be minimized. Data mining technique when combined with GIS contribute towards the new technology called spatial data mining. GIS was identified as the most appropriate tool for projecting the spatial distribution of different air and water quality parameters on the maps. In this way, the water quality problems could be identified and could be correlated with the activities taking place to interpret the reasons for deterioration of its quality.

The model using suitable data mining tools will analyze and compare the previous data and current data and project the future trend of data for upcoming years.

Data mining techniques were implemented to develop a Knowledge Database on the water and air quality of seven zones of Navi Mumbai viz. Nerul, Vashi, Belapur, Airoli, Ghansoli, Turbhe and Koparkhairne. The results at various stages of work are projected on maps using GIS.

This model will enable to develop a technique to get predictions as well as alerts for CIDCO (City and Industrial Development Corporation) or NMMC (Navi Mumbai Municipal Corporation) (Ministry of Health) to prevent further health victims by testing the air and water parameters and help them take remedial actions towards the health problems and also help them to preserve the existing stage of the air and water quality.

The prediction results demonstrated on seasonal health (Upper Respiratory Tract Infection) data for Turbhe and Vashi area revealed that that along with non-linear regression method, even additive regression method is also suitable. Hence both the methods can be applied on such type of data to get the accurate prediction.

The overall line of action of the research (right from the data-mining tool selection to model the air and water quality prediction) is described. Next, the comparison of the results of the four prediction algorithms applied on air quality data as well as on health Upper Respiratory Tract Infection (URTI) data and finally, the significance of practical demonstration of this research is discussed. Conclusions and future enhancement of this research has been presented at the end.

1.2 PURPOSE OF THE RESEARCH

This study is intended to demonstrate the comparison on human health due to air pollutants in industrial and residential zones. Association between air quality parameters and its effect on human health using various mining algorithms has been demonstrated as a case study.
The framework suggested entitles the researcher to design a model which is scalable, flexible, cost effective, and secure. It will work successfully for similar data or can be modified as per the requirements. It targets to suggest that data mining technique is the optimal solution to design and develop a user friendly interface that will not only be easy to access but will provide the necessary functionality also. The survey and analysis of pollution pattern on human beings using GIS can become further useful to study the impact and air pollutants on human health which is the main aim of this research.

This study will help

- Society and sensitize people about their surrounding environment.
- In implementing the data mining for assessment of environmental impact of pollutants.
- In predicting the future trend of data as an outcome depending on the previous assessment details.

1.3 BACKGROUND AND CONTEXT OVERVIEW

Earlier work by Dr. B.S. Mahajan was carried out through Health And Environment Action Based Learning (HEAL) project undertaken by Homi Bhabha Centre for Science Education (HBCSE), a branch of Tata Institute of Fundamental Research (TIFR), Mumbai. This project was carried out with the assistance of National Service scheme (NSS) cell of University of Mumbai. This project was executed under the guidance and monitoring of the author and trained teachers of the South Indian Education Society (SIES), Nerul College.

1.4 SCOPE OF THE RESEARCH:

1.4.1 Geographical Scope:

As environment is a very complex phenomenon, hence I have restricted my study for Air and Water parameters for Navi Mumbai area for all seven zones.

1.4.2 Conceptual Scope:

- Improved public health technical support for the management of the provision of the State and district health missions.
- Data collection, evaluation and evidence based planning, monitoring and supervision for strengthening the capacity for review.
There are certain issues and challenges of applying suitable actions for minimizing the hazardous parameters and efforts toward the perseverance of water body resulting in improved public health.

A lot of basic issues which affect the usability of WMS to support the embedding of data from all NGO’s or Government agencies providing health related data.

1.4.3 Analytical Scope:
As per the need of study for evaluation, assessment and interpretation of the data, GIS analysis is required. Also by using Data mining regression or decision tree, predictions can be drawn but subject to the availability of data for that particular area.

1.5 OBJECTIVES

- To suggest that GIS and data mining technology can be used to design and develop a scalable, flexible, cost effective, interactive, user friendly and secure model to assess an impact of air and water pollutants on human health.
- To Design, Develop and implement a model with an ability to integrate spatial data which will also analyze and manage the same, and also able to answer all queries related to locations well as health.
- To build a theoretical framework casing the diverse environmental and health domains in environmental monitoring system.
- To Survey the different techniques of data mining that can be applied in the environmental domains.
- To identify the classification (prediction) techniques which can be applied to appropriate environmental domains?
- To Implement the algorithms on the identified areas and analyze their accuracy and efficiency.
- To equip the decision makers with a handy tool to predict the trend of data as an outcome depending on the previous assessment details and take necessary remedial actions in the matter.
1.6 PROBLEM STATEMENT

To develop a user friendly, web based, scalable, secure, economic model which will help the facilitator to bridge the gap between the environmental data and its related health data. It will also help to analyze the results and gives the graphical representation of the analyzed data and also shows its impact on human health.

1.7 BENEFICIARIES

- Health Practitioners
- Policy makers (CIDCO,NMMC)
- Large corporate houses implementing Corporate Social Responsibility Initiatives (CSR)
- NGO’s concerned with and working on human health and environment
- Any environment conscious individual
- NMMC Health Department

1.8 CASE STUDY

ZONES STUDIED FOR WATER AND AIR

Select Zone

- NERUL
- YASHI
- BELAPUR
- TURBHE
- KOPARKHAIRNE
- GANSOLI
- AIROLI

Figure Error! No text of specified style in document.1 : Studied Zones of the Research area
The present study was confined to Navi Mumbai (19° 2’ 2.20” N, 73° 0’ 43.71”E). This city lies across the Thane creek, north east of Mumbai and flanked by the Thane creek waters on its west, south-west and north-west contours.

Navi Mumbai was developed as a planned city by City and Development Corporation of Maharashtra (CIDCO). There are 6 nodes and approximately 30-50 sectors in each node.

In several nodes the city has fresh water lakes (ponds). The Navi Mumbai Municipal Corporation (NMMC) and CIDCO have also impounded creek water creating a series of water holding ponds.

The present case study was confined to 7 different zones of Navi Mumbai which are: Nerul, Vashi, Belapur, Airoli, Ghansoli, Turbhe and Koparkhairne. Each zone with more than two water bodies and at-least one air station. After collecting the air and water pollutants data from these different zones at uniform interval of time, we applied data mining techniques for fining the missing data values and studied the trend of the data also predicted its future trend. By storing the rules in another database we could able to find out the diseases caused by these air and water pollutants from the same area. Also we are able to find out the upcoming number of health victims due to these reasons.

1.9 RESEARCH METHODOLOGY

The study of this project is based on combination of explorative and experimental research. The explorative part involves exploring the possibility of data mining and GIS technology to design a cost effective, scalable, secure, user friendly model which will help to assess environmental (i.e. air and water pollutants) impact on human health. The experimental part involves exploring the possibility of implementing data mining regression analysis tool to achieve the objective. This concept will be applied to design environmental impact assessment model and study the benefits of implementing this system.

1.10 DATA COLLECTION (SOURCES)

**Primary Data**

- HEAL data (Air & Water)
- Health Survey
Secondary Data

- Proprietary Data from NMMC for health, air and water
- Output of GIS i.e. analyzed data
- Books
- Documents published and unpublished
- Referred Journals and Research papers
- Internet

The source of the primary data collected as a result of experimental process of the HEAL project. Secondary data for Air and water is collected from the NMMC environmental labs for last 7 years. Also environmental status reports for past few years were studied to get related data. Health data is collected from NMMC health department and NMMC Environmental Status Report (ESR). Also Maharashtra Pollution Control Board data is also collected to study its trend, and find its impact on human health.

Then by applying data mining tool data is preprocessed and analyzed. Then it will be connected to GIS to get answers all the health related queries through the interactive map. By studying the previous and current data the future trend can be found out by using time series or regression analysis. The neighborhood analysis to find out the information about nearby objects can be done as and when required either through GIS or through Data mining neighborhood analysis.

1.11 FEASIBILITY STUDY

1.11.1 Technical Feasibility

The technology required for the development of this system is available easily that is JAVA (Swing) and SQL SERVER 2005, ArcGIS/GRAM++, Data Mining software Weka. It is made to ensure that technology required for developing the system is available. The model proposed here has the capacity to hold the required data. The system can be expanded in future if needed.

1.11.2 Economic Feasibility

Once the required software and hardware requirements get fulfilled, there is no need for user of our system to spend for any additional overhead. For the user, the system will be economical
feasible in following aspects. The system will reduce a lot of paper work hence the cost will be reduced. There will be no training cost as the system has very user friendly GUI.

### 1.11.3 Operational Feasibility

- Total working hours required will be less.
- Reduced manpower.
- Real time reports are generated.
- More secured because of password protection.

The system is operationally feasible because of above tested benefits.

### 1.12 Features of the Proposed System

1. **Analysis:** The existing system can take decisions based on the past data and with help of trend analysis we can obtain the forecasted output.

2. **Cost Effective:** As the software is open source so it is not costly.

3. **Reduction of Paper Work:** Paper work is reduced as all the records are stored on databases in the computer, thus reducing manpower and time.

4. **High Performance:** Traditionally for data retrieval, one has to go through piles of files and papers. But with this new system data retrieval is just a click way.

5. **Minimized Redundancy:** Whenever some data is updated in the traditional system, it leads to more paperwork. Thus a proper system tied to automation can minimize the problem of redundancy.

6. **Interactive:** The records are represented in the form of maps and graphs. So it becomes more attractive.

7. **Security:** Security is maintained through registration and login facility.

8. **User Friendly:** The system is very interactive and has easy access to the data. Read only content is provided with a simple navigation and map.

9. **Compatible:** It is compatible with all the operating systems.

10. **Administrative Control:** Administrator can have access to the database.
11. **Scalable:** More modules can be added in future.

12. **Expandable:** Our system is based on few zones of Navi Mumbai but it can be expandable for other zones, cities and states.

### 1.13 CONCLUSION

This is an initiative to design and develop a tool which scalable, economic, robust, flexible, user friendly and is useful to predict the future air and water quality data.

It will also help the facilitators, policy makers, health practitioners to check its impact on human health and take the preventive measures.

The outcomes have recognized the correlation between the air quality and the health of the human beings related to air quality.

It is hoped that this project, can help in the efforts to prevent further deterioration of our environment and reduce the related disease burden. It will also help in fast retrieval of the data as this model is using very strong model developed in POSTGRESQL as back end and Java (Swing) as front end as well as WEKA software for future forecasting. Such type of water and air monitoring can be followed in various area at different scales of operation has become a strong possibility today. Hence if same type of data is collected which will fit in this model, then this model can be executed.