International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) e-ISSN: 2395-5325
Volume3, Issue 2 (February '2017)

IMPACT OF ENVIRONMENTAL POLLUTION: A DATA MINING APPROACH

S. Benitta Roseline,
PG Student,
Department of Computer Science,
Jayaraj Annapckiam College for Women (Autonomous),
Periyakulam,Tamilnadu,India.

S. Sophia Rani,
Assistant Professor,
Department of Computer Science,
Jayaraj Annapckiam College for Women (Autonomous),
Periyakulam,Tamilnadu,India.

S. Soundarya Jothi,
PG Student,
Department of Computer Science,
Jayaraj Annapckiam College for Women (Autonomous),
Periyakulam,Tamilnadu,India.

Abstract: Environmental pollution is one of the phenomena that and have gained international attention. This study was born out of the inquisitives to verify the problems associated with the refuse disposal and how they affect the personal and environmental health of inhabitants. Environment concerns have become a major development issue in Nigeria as can be seen in the establish mental agency and its equivalent at state levels and in all the local governments of imo state. The growing menace of air pollution is one of the crucial environmental issues in modern area across the world. This is silent killer of the physical as well as mental health of human being. Ten research questions were formulated for the study. The data collection using percentages. Findings from the study were discussed based on the finding suggestions were made on throw to improve on environmental refuse disposal for good personal and environmental health. The suggestions were made not only the local government area but also to the society, parents especially women, school and wards of every household.

Keywords: Environmental pollution ,Data mining ,Apriori Algorithm

I.INTRODUCTION

Environmental Pollution is an international journal that seeks to publish papers that report results from original, novel research that addresses significant environmental pollution issues and problems and contribute new knowledge to science. The editors welcome high quality papers where the pollutants are clearly defined and measured and can be directly related to biological, ecological, and human health effects. This includes air, water, and soil pollution and climate change. New techniques for the study and measurement of pollutants are also encouraged. Papers, such as Meta analyses, that report findings from re-examination and interpretation of existing data are welcome Modeling papers are also welcome Critical review papers are also of interest as are letters to the editor. The editors do not wish to publish papers that describe results from routine surveys and monitoring programs that are primarily of local interest. Descriptions of well-known pollutants, such as legacy pollutants, in yet another location are not of interest. Papers about sewage, waste and wastewater treatment and management, agronomy, remediation, biomonitoring, bioremediation and phytoremediation are not acceptable.

Present Condition: The progress of industrialization has been accompanied by a steady rise in emission levels. In China, for example, SO2 emissions surged from 15.23 million tons in 1985 to 17.95 million tons in 1993. However, there have also been cases of substantial abatement of pollution once industrialization has progressed to a certain extent. In the Republic of Korea, for example, SO2 emissions improved considerably in the late 1980s. An analogous improvement was achieved by Japan in the 1970s; yearly SO2 emissions, which probably topped 4.8 million tons in the late 1960s, have been reduced to about 1 million tons since the 1980s. These cases show that air and water pollution from industrial processes can be largely corrected if the proper technical measures are taken. However, there also exist problems for which substantial improvement has not been achieved even in countries possessing sophisticated technology, as well as problems which surface along with the emergence of new technology. In Japan, groundwater has been polluted by chemical substances (such as trichloroethylene) used in the fabrication of semiconductors. Similarly, many Asian countries are grappling with problems associated with the storage and disposal of large quantities of hazardous chemical substances used in semiconductor fabrication. A great risk is also posed by marine pollution by oil tankers in the Straits of Malacca and other bodies of water in Asia. Furthermore, industrialization is being linked to the spread of mass-production and consumption-oriented lifestyles throughout Asia. Burial of the vast quantities of resulting waste in inland and coastal areas is having an adverse impact on the natural environment. In some cases, groundwater is being polluted by waste stored in the open.

Significance of the study: This section, often referred to as the "rationale" is crucial, because it is one place in which the researcher tries to convince an audience that the research worth doing. It should establish why the audience should want to read on. Many researchers believe that the modern world faces enormous environmental problems that have been created by humanity's interference with nature's cycles. Managing changes in climate, population growth and water resources will require professional's expertise, if we intend to achieve sustainable development. Knowledge about various kinds of pollution and their effects will be increasingly

important for planning and management in the future. If you are interested in an education that provides expertise in environmental protection work in connection with the contamination of soil water and air as well as other environmental issues, such as noise, waste, indoor climate and radiation, you should consider enrolling in the Bachelor's degree in the environmental pollution and the environment. There is currently very few other study programmers in Norway that cover these important social tasks.

Objectives of the study: The advance pollution prevention by maintaining and improving opportunities for exchanging ideas and facilitating coordination of efforts. The environmental pollution has many objectives.

- Maintain and improve annual NPPR conferences as important formous for exchanging ideas and meeting service needs.
- Increase access to member network expertise by facilitating communication through all available media.
- Enhance effectiveness of workgroups.
- Support and promote a national pollution prevention activities of regional roundtables and academic pollution prevention initiatives.

Null hypotheses: Pollution is an undesirable change in the physical, chemical, or biological characteristics of the natural environment, brought about by man's activities. It may be harmful to human or nonhuman life. Pollution may affect the soil, rivers, seas, or the atmosphere. There are two main classes of pollutants: those that are biodegradable (e.g. sewage), i.e. can be rendered harmless by natural processes and need therefore cause no permanent harm if adequately dispersed or treated; and those that are non-biodegradable (e.g. heavy metals (such as lead) in industrial effluents and DDT and other chlorinated hydrocarbons used as pesticides), which eventually accumulate in the environment and may be concentrated in food chains. Other forms of pollution in the environment include noise (e.g. from jet aircraft, traffic, and industrial processes) and thermal pollution (e.g. the release of excessive waste heat into lakes or rivers causing harm to wildlife). Recent pollution problems include the disposal of radioactive waste; acid rain; photochemical smog; increasing levels of human waste; high levels of carbon dioxide and other greenhouse gases in the atmosphere (see greenhouse effect); damage to the ozone layer by nitrogen oxides, chlorofluorocarbons (CFCs), and halons; and pollution of inland waters by agricultural fertilizers and sewage effluent, causing eutrophication. Attempts to contain or prevent pollution include strict regulations concerning factory emissions, the use of smokeless fuels, the banning of certain pesticides, the increasing use of lead-free petrol, restrictions on the use of chlorofluorocarbons, and the introduction, in some countries, of catalytic converters to cut pollutants in car exhausts.

II.APRIORI ALGORITHM

Apriori algorithm is a fundamental algorithm mining association rule. It contains two processes: Detect all frequent item sets by scanning db. Form strong association rules in the frequent item sets. Process one needs to scan DB several times, which consumes a lot of time and space. As a result, what needs to be improved is the mining competency of frequent group of things in DB. Apriori algorithm is a significant algorithm for mining frequent item sets for Boolean association rules. Apriori algorithm is formed by Agrawal and Srikantin 1994. It is the most fundamental and important algorithm for mining frequent item sets. Apriori is used to detect all frequent item sets in a provided database db. The keynote of Apriori algorithms is to form multiple passes over the database. It employs an repetitive approach called as a breadth-first search (level-wise search).

Key Concepts:

- **Frequent Item sets:** The item sets which minimum help (denoted by li for ith –item sets), Apriori property: any subgroup of frequent things must be frequent.
- **Join Operation:** to detect lk, a group of candidate k-group of things is developed by adding lk-l with itself.

How Apriori Works?

- **Find All Frequent Item sets.**
- **Get Frequent Things:** Things whose occurrence in database is more than or equal to the minimum help threshold.
- **Frequent Item sets**: Develop candidates from frequent things. Prune the results to detect the frequent item sets. Develop strong association rules from frequent. Rules which satisfy the minimum support and minimum confidence threshold.
- **Association Rules:**Association rule of data mining involves picking out the unknown inter-dependence of the data and finding out the rules between those items [3]. Agrawal introduced association rules for point of sale (POS) Systems in supermarkets. A rule is defined as an implication of the form A=>B. The left-hand side of the rule is called as antecedent. The right-hand side of the rule is called as consequent.
- **Support:** I={i1,i2,i3,……,im} is a collection of items. T be a collection of transactions associated with the items. Every transaction has an identifier TID [6]. Association rule A=>B is such that A C1,B C1.A is called as premise and B is called as Conclusion. The support, S, a is defined as the proportion of transactions in the data set which contains the item sets
- **Support(X=>Y)=Support(XUY)=P(XUY).**
- **Confidence:** The confidence is defined as a conditional probability Confidence(X=>Y) = Support (XUY)/Support(X) =P(Y/X).
  - **Lift:** is the ratio of the probability that L and R occur together to the multiple of the two individual probabilities for L and R, lift=pr (L, R)/Pr (L).Pr(R).

2.2 Statistical Techniques Used

Data sources and methodology

Target population : This survey covers all peoples in Theni District

Instrument design: The data has been collected from Theni district peoples. The items on the questionnaire have remained unchanged for several years. However should

modifications become necessary, proposed changes would go through a review committee, and a field test with respondents and data users to ensure its relevancy.

Sampling: This survey is a census with a cross-sectional design. Data are collected for all units of the target population, therefore no sampling is done.

Error detection: There are edits build into the data capture application to check the entered data for unusual values as well as to check for logical inconsistencies. Whenever an edit fails ,the interviewer is prompted to correct the information(with the help of the respondent when necessary).For most edit failures the interviewer has the ability to override the edit failure if necessary.

Imputation: A 100% response rate is attained; therefore imputation is not necessary.

Quality evaluation: Prior to the data release, combined survey results are analyzed for comparability; in general this includes a detailed review of individual responses, general economic condition, and historical trends. The data is examined at a macro level to ensure that the long-term trends make sense when compared to publicly available information in media reports, company press releases, etc. Large fluctuation in year-over-year sales and inventories are analyzed to determine if they are in error or if they accurately reflect retails activity. Subjected matter officers follow up with the company to confirm the data and to document reasons for large fluctuations in sales or inventories.

Disclosure control: Releasing any data would information obtained under the statically Act that relates to any identifiable person, business or organization without the prior knowledge or the consent in writing of that person, business, or organization. Various confidentiality rules are applied to all data that are released or published to prevent the publication or disclosure of any information deemed confidential. If necessary, data are suppressed to prevent direct or residual disclosure of identifiable data.

Revisions and seasonal adjustment: Revisions in the raw data are required to correct known non-sampling errors. These normally include replacing imputed data with reported data, corrections to previously reported data, and estimates for new births that were not known at the time of the original estimates. Raw data are revised, on a monthly basis, for the monthly immediately prior to the current reference month being published. The purpose is to correct any significant problems that have been found that apply for an extended period. The actual period of revision depends on the nature of the problem identified, but rarely exceeds three years.

III. ANALYSIS OF DATA

Data mining definition: Data Mining is the process of analyzing data from different perspective and summarizing it into useful information-information that can be used to increase revenue, cuts costs, or both. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost.

Data, Information and Knowledge:

Data: Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases, This includes:

- Operational or transactional data such as sales, cost, inventory, payroll, and accounting
- Nonoperational data, such as industry sales, forecast data, and macro economic data
- Meta data- data about the data itself; such as logical database design or data dictionary definitions.

Information: The patterns, associations. Or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

Knowledge: Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior . Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Data Mining Process: Data Mining is primarily used today by companies with a strong consumer focus-retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price. Product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And. It enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data. With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

How does data mining work?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyze relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- **Classes:** Stored data is used to locate data in predetermined groups. For example. A restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **Clusters:** Data items are grouped according to logical relationships or consumer preferences. For example, data

can be mined to identify market segments or consumer affinities.

- **Associations:** Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **Sequential patterns:** Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data Mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analyst and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

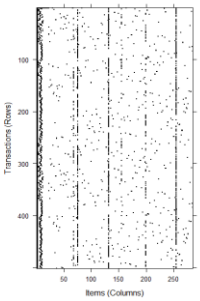
Different levels of analysis are available:

- **Artificial neural networks:** Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Genetic algorithms:** Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- **Decision trees:** Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include classification and Regression Tree (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rule that you can apply to a new (unclassified) dataset to predict which record will have a given outcome. CART segment a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
- **Nearest neighbor method:** A technique that classifies each record in a dataset based on a combination of the classes of the k record most similar to it in a historical dataset. Sometimes called the k-nearest neighbor technique.
- **Rule induction:** The extraction of useful if then rules from data based on statistical significance.
- **Data visualization:** The visual interpretation of complex relationship in multidimensional data. Graphics tools are used to illustrate data relationships.

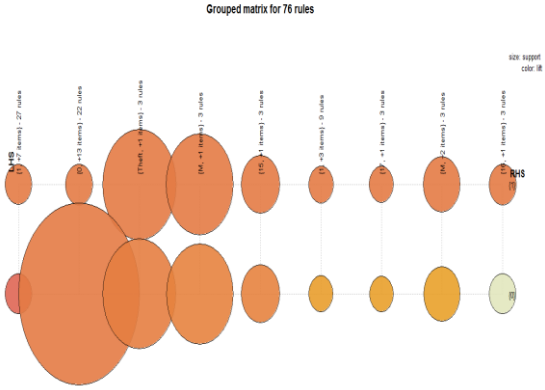
Experiment Results:

```
>                                                beni<-
readLines("C:/Users/MURUGAN/Desktop/pollution1.csv"
)
>                    local({pkg                    <-
select.list(sort(.packages(all.available           =
TRUE)),graphics=TRUE)
+             if(nchar(pkg))             library(pkg,
character.only=TRUE)})
Loading required package: Matrix
Attaching package: 'arules'
The    following    objects    are    masked    from
'package:base':
    abbreviate, write
Warning message:
package 'arules' was built under R version 3.2.4
> library(arules)
> patterns=random.patterns(nItems=150);
>
trans=random.transactions(nItems=150,nTrans=150,met
hod="agrawal",patterns=patterns);
>  summary(patterns)
set of 2000 itemsets
most frequent items:
 item21  item71  item89  item62  item39 (Other)
    257     225     175     174     169    7097
element          (itemset/transaction)          length
distribution:sizes
  1   2   3   4   5   6   7   8   9  10  11
 82 298 425 478 339 206 109  39  16   6   2
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   4.000   4.048   5.000  11.000
summary of quality measures:
    pWeights            pCorrupts
 Min.   :8.400e-08   Min.   :0.0000
 1st Qu.:1.565e-04   1st Qu.:0.2768
 Median :3.519e-04   Median :0.4961
 Mean   :5.000e-04   Mean   :0.4936
 3rd Qu.:6.913e-04   3rd Qu.:0.7090
 Max.   :4.293e-03   Max.   :1.0000
includes transaction ID lists: FALSE
>          groupfsets=eclat(trans,          parameter=
list(support=0.05), control=list(verbose=FALSE));
> fsets=eclat(trans, parameter= list(support=0.05),
control=list(verbose=FALSE));
> singleItems=fsets[size(items(fsets))==1];
> library(utils)
> library(methods)
> library(datasets)
> library(base)
> singleItems=fsets[size(items(fsets))==1]
> singleSupport=quality(singleItems)$support
>
names(singleSupport)=unlist(LIST(items(singleItems)
,decode=FALSE));
>  head(singleSupport, n=5);
       21        71        89       116        62
0.4000000 0.2733333 0.2333333 0.2000000 0.2133333
> itemsetList=LIST(items(fsets),decode=FALSE);
>
allConfidence=quality(fsets)$support/sapply(itemset
List,
function(x)+max(singleSupport[as.character(x)]));
>
quality(fsets)=cbind(quality(fsets),allConfidence);
> summary(fsets)
set of 139 itemsets
most frequent items:
 item21  item71 item116  item29  item89 (Other)
     33      17       9       7       7     131
element          (itemset/transaction)          length
distribution:sizes
 1  2  3
77 59  3
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.468   2.000   3.000
summary of quality measures:
    support         allConfidence
 Min.   :0.05333   Min.   :0.1333
 1st Qu.:0.06000   1st Qu.:0.2439
 Median :0.08000   Median :1.0000
 Mean   :0.09319   Mean   :0.6542
 3rd Qu.:0.11333   3rd Qu.:1.0000
 Max.   :0.40000   Max.   :1.0000
includes transaction ID lists: FALSE
mining info:
  data ntransactions support
 trans           150    0.05
```

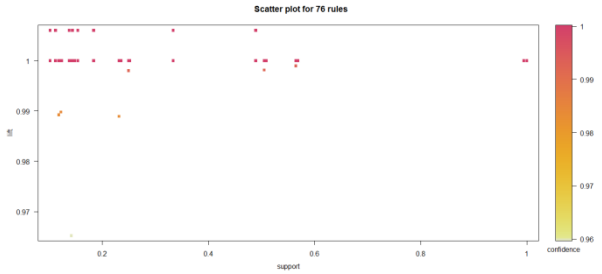
```
> library(arules)
>     tr<-   read.transactions("Div.csv",   format   =
"basket", sep=',', rm.duplicates=TRUE)
distribution of transactions with duplicates:
items
 19  20
498   2
> rules<-apriori(tr)
Apriori
Parameter specification:
 confidence minval smax arem  aval originalSupport
support minlen maxlen target
        0.8    0.1    1 none FALSE            TRUE
0.1      1     10  rules
   ext
 FALSE
Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE
Absolute minimum support count: 50
set item appearances ...[0 item(s)] done [0.00s].
set     transactions     ...[284     item(s),     501
transaction(s)] done [0.00s].
sorting  and  recoding  items  ...  [14  item(s)]  done
[0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 done [0.00s].
writing ... [76 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
> image(tr)
```



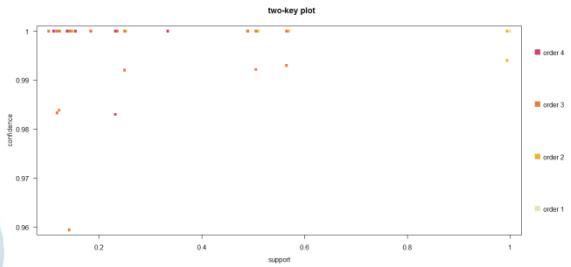

```
> library(arulesViz)
Loading required package: grid
Warning messages:
1: package 'arulesViz' was built under R version
3.2.4
2: replacing previous import by 'utils::head' when
loading 'arulesViz'
>                   local({pkg                    <-
select.list(sort(.packages(all.available          =
TRUE)),graphics=TRUE)
+            if(nchar(pkg))            library(pkg,
character.only=TRUE)})
> plot(rules,method="grouped")
> plot(rules,method="grouped")
```



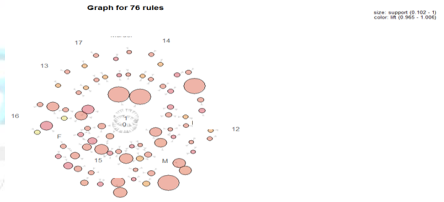

```
head(quality(rules));
    support confidence     lift
1 0.9940120  0.9940120 1.000000
2 1.0000000  1.0000000 1.000000
3 0.1017964  1.0000000 1.006024
4 0.1017964  1.0000000 1.000000
5 0.1177645  0.9833333 0.989257
6 0.1197605  1.0000000 1.000000
>
plot(rules,measure=c("support","lift"),shading="con
fidence");
```



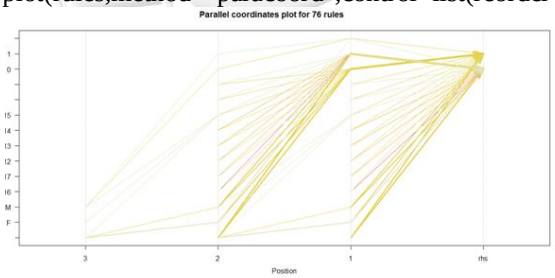

```
plot(rules,shading="order",control=list(main="two-
key plot"));
```




```
plot(rules,method="graph")
```




```
plot(rules,method="paracoord",control=list(reorder=TRUE)
```



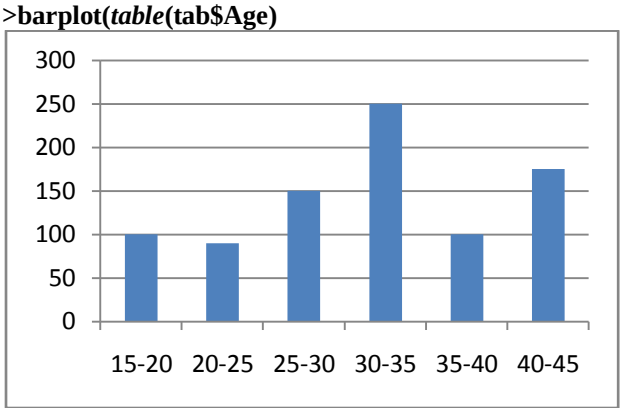

```
> plot(density(beni$pollution1))
```

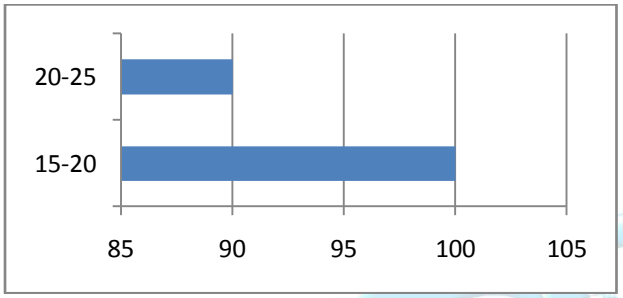
IV. INTERPRATIONS, RECOMMENDATIONS AND SUGGECTIONS

Statement of the problem: The National Pollution Prevention Roundable is a national forum that promotes the development, implementation, and evaluation of efforts to avoid, eliminate, or reduce waste generated to air, land, and water. The sustainable and efficient use of energy, meterials and resources is vital to the protection and enhancement of human health and the environment, and the conservation of natural resources. These efforts are integral to accelerate the shift towards sustainable consumption and production to

promote environmental responsible social and economic development.

>barplot(*table*(tab$Age)





Recommandations: Specific actions recommended for the control of sediment, fertilizer and pesticide impacts on water quality have been enumerated in each respective chapter. In this concluding chapter are discussed a selection of issues pertaining to water quality which have over-arching implications for agriculture. Some, such as internalization of all environmental costs, represent a significant challenge to the way in which agricultural agencies pursue agricultural practices. Others reflect the interaction between agriculture and other national development priorities and reflect public policy issues that are of national significance.

- The environmental impacts on water resources [in general] caused by agricultural activities cannot be disassociated from the agricultural impacts in production areas themselves. They require monitoring, and preventive measures should always be systemically integrated.
- Data and information generated should be properly treated in the sense of disseminating them as much as possible in order to heighten awareness and mobilization of the public sector and of society with respect to agriculture's impact on the environment.
- Attempts should be made to exchange information and to pursue horizontal cooperation among countries, in order to promote the exchange of information and experiences.
- Besides treating water quality related problems, there is evidence of other problems generated by conflicts in use, particularly the need to integrate quality management with the quantity of water within a comprehensive, decentralized and participatory management system, reconciling regional development with environmental protection.
- The cooperation of fiscalizing and monitoring agencies dealing with agrotoxic substances is urgent, with the capacity for control structures, seeking the development of biological indicators (enzymes, AMES test, biotesting, bioindicators) of residues and of anatomy and pathological damages caused by agrotoxic substances.

V.CONCLUSION

It appears that polluted environment is global an issue and world community would bear worst results more as they already faced. As effective response to pollution is largely based on human appraisal of the problem (Kromm, 1973) and pollution control program evolves as a nationwide fixed cost-sharing effort relying upon voluntary participation (Sharp & Bromley, 1979).Education, research, and advocacy, are lacking in the region as preventive strategy for pollution. Journal of Arts, Science & Commerce International Refereed Research Journal Vol.– II, Issue –2,April 2011 281(Fitzgerald, 1998) especially in Asia. At present the adoption of environmental auditing in any economic sector is voluntary but future legislation could well make it mandatory (Goodall, 1995) and still time available to use technology and information for environmental health decision. Policymakers in developing countries need to design programs, set standards, and take action to mitigate adverse health effects of air pollution. Healthy people mean human resources are the main object of any successful business or country. These societal beneficial efforts need to carefully adapt available knowledge from other settings, keeping in mind the differences in pollutant mixtures, concentration levels, exposure patterns, and various underlying population characteristics.

VI.REFERENCES

[1] Ashraf, M. A., Maah, M. J., Yusoff, I. & Mehmood, K. (2010). Effects of Polluted Water Irrigation on Environment and Health of People in Jamber, District Kasur, Pakistan, International Journal of Basic & Applied Sciences, 10(3), pp. 37-57.

[2] Avdeev, O. & Korchagin, P. (1994). Organization and Implementation of Contaminated Waste Neutralization in the Ukraine - National Report II, Central. European Journal of Public Health, 2(suppl), pp. 51-52.ss

[3] Blaxill, M. F. (2004). What's going on? The Question of Time Trends in Autism. Public Health Reports, 119(6), pp. 536-551.

[4] Brauer, M., Hoek, G., Smith, H. A., de Jongste, J. C., Gerritsen, J. & Postma, D. S. (2007). Air Pollution and Development of Asthma, Allergy and Infections in a Birth Cohort, European Society for Clinical Respiratory Physiology, 29(5), pp. 879-888.

[5] British Airways, (1993), Annual Environment Report, London: British Airways plc, Environment Branch, Heathrow.