International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Spatio-Temporal Clustering for Environmental Data: India

Mariyam Kidwai[1], Garima Srivastava[2]

[1, 2]Amity School of Engineering and Technology (ASET Dept.), Amity University, Lucknow Campus, Uttar Pradesh, India

Abstract: ***Data mining is the mathematical approach of exploring new information by assessing definitely known facts and figures. The paper applies this approach to the real-world environmental data: The Greenhouse Gas Emissions for the country India. The greenhouse gas effect is a gradual phenomenon of increase in the average temperature of the earth's atmosphere due to the raised emissions and concentrations of greenhouse gases. Global Warming, which is a topical issue, has led to extreme global climate change posing as a mammoth threat to the foods security situation in India with recurring and severe droughts and ravaging floods engulfing the arable land. [3] In this paper, a novel unsupervised learning phenomenon has been used in finding out the high and low Greenhouse gas emissions' sectors in India comparing with China and Indonesia (top three highly populated Asian countries) for fair analysis. The three countries (incorporating space) and their consecutive 15 years data (incorporating time) result into large spatio-temporal data sets that has raised the need of performing spatio-temporal clustering on the data sets. The CRISP-DM methodology, K-means clustering algorithm and Weka tool has been used to design and develop the model which can be used to analyze environmental data by governmental authorities when decisions on such data are to be made and will also provide deeper in-sight of the data, thus, contributing towards the widespread efforts to reduce the greenhouse gas emissions to mitigate weather changes and promote cleaner energy sources.***

Keywords: Asian countries, CRISP-DM Methodology, Data Mining, Greenhouse Gas Emissions, K-means clustering algorithm, Spatio-Temporal Clustering, Unsupervised learning, Weka

1. Introduction

Data mining is the mathematical approach of assessing large pre-existing data sets in order to extract new information. Due to the revolutionary technological advancements and the world of Internet growing wider, the amount of data is also growing much faster leading to the terms Terabyte (10^12), Petabyte (10^15), Exabyte (10^18), Zettabyte (10^21) and Yottabyte (10^24) to describe the amount of Big data. [1] The aim of this paper is to apply data mining approach to the real-world big data of the environment: The Greenhouse Gas Emissions for the country India. Carbon dioxide (CO2), methane (CH4), nitrous oxide (N2O), etc. popularly known as the Greenhouse gases act like a blanket around the earth, trapping excessive heat that would have otherwise escaped into the space which in turn is gradually increasing the average temperature of the earth's atmosphere. This effect is termed as the Greenhouse Gas Effect resulting into the Global Environmental Change. The economy of India is the seventh largest economy in the world and is classified into three sectors. According to the World Bank data, the percentage contribution of three sectors to the total GDP of India in the year 2012 is:

Table 1: Sectors and their contributions to India's economy

Sectors	*% Contribution*
Agriculture Sector	18.7
Industry Sector	31.7
Service Sector	49.6

Global warming in this century is the watchword, with its adverse impacts as of now being conveyed to lime-light by the repeating occasions of enormous floods, annihilating droughts and ravaging cyclones all through the globe. According to some of the World Bank findings related to Global Warming impacts for India, the devastating impacts of Global Warming to the different sectors of India include extreme heat, changing rainfall patterns, droughts, exploitation of groundwater, more than 60% of India's agriculture is rain-fed, making the country highly dependent on groundwater; even without climate change, 15% of India's groundwater resources are overexploited, glacier melt, sea level rise, agriculture and food insecurities, seasonal water scarcity, rising temperatures, and intrusion of sea water would threaten crop yields, jeopardizing the country's food security, energy insecurity, water insecurity, health, migration and conflict, thus, affecting the economic growth of India. [2] India is as of now a catastrophe inclined zone, with the insights of 27 out of 35 states being calamity inclined, with most debacles being water related. [3] The process of global warming has led to an increase in the frequency and intensity of these climatic catastrophes. [3] The Indian economy is considered as one of the fastest developing economies. Nonetheless, the nation is tormented by the climatic calamities that keep on wreaking ruin on its economy. Therefore, disregarding the leaping economic advancement, majority of India's population keep on living in poverty, with lack of healthy sustenance and maladies consuming the society. Warming of the climate system is unequivocal. [4] A rise of 0.5 degree Celsius in winter temperatures could cause a 0.45 ton per hectare fall in India's wheat production. [4] In this paper, the greenhouse gas emissions from the different sectors in India over the past 15 years is analyzed, pre- processed, normalized and clustered to give a model that would prove to be beneficial to the environment-related authorities while coping up with the global warming borne environmental risks and problems that are heading towards us. India is the second largest populous country in the world. For comparison, we have also taken China and Indonesia that are among the highly populated countries in the world for fair computations. Lastly, the changes in the greenhouse gas emissions that took place over time among the countries are examined. In this research, the greenhouse gas emissions in terms of the major pollutants give the high and low emission sectors. [5]

2. Design Methodology

A CRISP-DM

Cross Industry Standard Procedure for Data Mining (CRISP-DM) methodology as the name is, is a standard procedure for designing data mining based applications. The methodology is preferred by a majority of data miners since it has well defined steps and structured approach to perform the tasks.

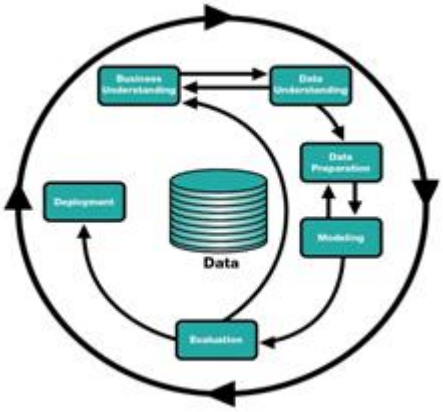


Figure 1: CRISP-DM Model [6]

This methodology is suitable to meet the objectives of this research. Also, the steps of CRISP-DM are very much similar to those of the steps of the WEKA visualization tool used in this research. The various steps of this methodology are shown in the above figure.

1) *Business Understanding*: In this phase, we need to find out and understand our field of interest. Then, we have to do some research work in order to find out the particular topic from our field of interest that has broader scopes to be mined. It also aims to ensure that the objectives of the research are clearly defined and understood.
2) *Data Understanding*: This phase aims at the basic understanding of the data to be used in the research, the quality and reliability of the data. In this phase, the exploration, collection, description, analysis and integration of data is carried out.
3) *Data preparation:* This phase generally includes selection of data, cleaning of the data, construction, and integration and formatting of the data. In data selection, the irrelevant attributes are filtered in order to obtain only those attributes that produce the optimal desired output. In data cleaning, all types of data inconsistencies are resolved. In data construction, new data is generated from the existing data if necessary. In data integration, separate data records are integrated. And in data formatting, the data is converted into the desired format in accordance to the techniques and tools to be used for computation and visualization.
4) *Data Modeling:* In the modeling phase, the prepared data is computed using a particular technique and also uploaded to the modeling tool in order to build the model of the reformed data.
5) *Evaluation:* In the evaluation phase, the model generated from the modeling phase is evaluated to get the data mining results.

B WEKA *Explorer*

The most important graphical user interface of *Waikato Environment for Knowledge Analysis (WEKA)* is the knowledge "*Explorer*". There are different panels featured by the Explorer interface namely:

1) *Pre-process:* In pre-processing, the data can be imported from the database, browsed and opened in particular formats (.csv, .arff, .names, .data etc.). Also, the data can be transformed by using filters in order to obtain desired attributes and instances.
2) *Classify:* We can apply different classification and regression algorithms to the data.
3) *Associate:* Association rule learners are accessible in this panel in order to detect relationships between the attributes of the datasets.
4) *Clusters:* Several types of clustering techniques (ex., Simple K-means, EM etc.) are provided in this panel.
5) *Select:* The most predictive attribute in the data can be identified with the help of the algorithms provided in this panel.
6) *Visualize:* In this panel, the scatter-plot matrix is shown where each individual scatter plot can be enlarged for further analysis by using the several selection operators.

C WEKA *KnowledgeFlow*

WEKA *KnowledgeFlow* is also a graphical user interface of Weka which provides a "data-flow" inspired interface to Weka by building the data-flow model of the entire processes involved in the mining of the data. The various components from the toolbar of the KnowledgeFlow are placed on the layout canvas and are connected together to produce a diagram similar to that of a knowledge flow for processing and analyzing data showing the pre-processing and mining techniques involved and visualization of the results produced. The KnowledgeFlow model represents the data preprocessing, modeling and evaluation phases of the CRISP-DM methodology. Evaluation, Visualization, Filtering, Classification and Data Source components are available in KnowledgeFlow along with a number of attributes to be selected accordingly for the formation of the data flow.

3. Implementation via K-Means, Spatio-Temporal Clusterings

The implementation of the research is explained in this section. Only the necessary and important steps taken, changes and decisions made during the use of CRISP-DM cycle, Weka Explorer and Weka KnowledgeFlow interfaces are documented while reaching to the final model.

1. Data Understanding: The data used in this research is acquired from the internationally recognized World Bank Organization which is free and publicly available data. The data comprises of 7 attributes: country, population, pollutant name, emissions, unit, year and sector. The following Table 2 gives an overview of each of the 7 attributes including their data types:

Table 2: Data Description

Attributes	Attribute description	Data-Types
Country	The name of the country	Nominal
Population	The population of the country	Numeric
Pollutant Name	The pollutant name represented in the record	Nominal
Emissions	The value of the emissions	Numeric
Unit	The unit that the emissions are measured in	Nominal
Year	The year in which the emission was recorded	Numeric
Sector	The unique sector code	Numeric

The emissions of the three pollutants CO2, CH4 and N2O are taken since they are considered to be highly responsible for Greenhouse Gas Effect. The three sectors namely: Industrial Activities/Production, Human Agricultural Activities and Biomass Burning and Livestock Management are considered in this research.

2. Data Preparation: The data selection, cleaning, construction, integration and formatting are carried out in this phase.

a) *Data Selection:*
- *Attribute Selection:* All the attributes shown in Table 2 are relevant, thus, all of them are included in this research.
- *Country Selection:* For the attribute country, India is the country which is focused in this research. Apart from that, China and Indonesia are also taken into consideration for comparing with India. For fair analysis, the countries with similar population size need to be taken into account and these three countries are among the top three highly populated Asian countries.
- *Sector Selection:* The culprit is greenhouse gases, notably Carbon dioxide, Methane and Nitrous oxide. These are accumulating to unprecedented levels in the atmosphere as a result of profligate burning of fossil fuels, industrial processes, farming activities and changing land use. [2] According to the reports from EDGAR and World Bank, CO2 is generally emitted due to the industrial activities, CH4 emissions take place from the human activities such as agriculture and N2O production mostly takes place from biomass burning and livestock management. Thus, the three sectors: 1) Industrial Activities/Production, 2) Human Agricultural Activities and 3) Biomass Burning and Livestock Management are taken into account for this research.

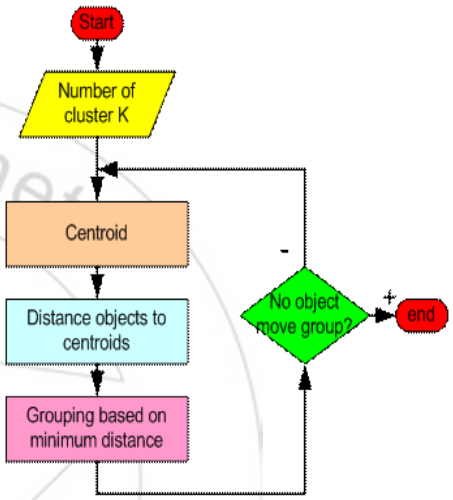
b) *Data Construction*: The 15 years data(from FY 1997-98 to FY 2011-12) is used, so three sets of 15 individual Excel sheets are produced, one for each year for each of the three countries.

c) *Formatting:* The Excel sheets are then converted into the .csv files since Comma Separated Value (.csv) extension files are supported in Weka.

d) *Filter:* All the .csv files are loaded into Weka Explorer and are passed through the unsupervised attribute filter "normalize" in order to obtain all the attribute values within the suitable range by using Scaling Factor, S=1.0 and Translation Factor, T=0.0.

e) Modeling: In this phase, the data, so prepared, is computed through the clearly defined, well specified and more interactive method of clustering provided by the Weka Explorer interface and then, the final model is prepared by using the several components from the toolbar of the Weka KnowledgeFlow interface.

f) Technique Selection: The clustering technique selected for this research is the unsupervised learning algorithm: the simple k-means clustering according to the requirement alongside the observation obtained through the study of several research papers and literature reviews. The requirement is we need to obtain an easy-to-understand visualization of the percentage of high and low emitters and the selected technique meets this objective. The simple k-means procedure is shown in the following:



Flowchart 1: k-means clustering [7]

which can be interpreted as:

1) Randomly select k number of clusters in the dataset (in this research, k=2 since two clusters: the "high emitters" and the "low emitters" need to be found.),
2) Allot the k points to the cluster centroids,
3) Allocate each point in the dataset to its nearest cluster center,
4) Move each cluster center to the mean of its assigned points,
5) Repeat 2-4 until convergence.

6) Building the Model: In the preprocess panel, all the three sets of 15 .csv files are opened in the Weka Explorer separately. The unsupervised attribute filter "normalize" is selected. The cluster panel is selected. All of the attributes except "Emissions" are ignored allowing for the simple k-means clustering algorithm to cluster the data based only on the value of "Emissions". The use of training set is selected for the assurance of the clustering algorithm computing records accurately. The algorithm in Weka Explorer uses Euclidean distance for measuring the distance between two points. The cluster centroids for each country are noted down into Excel sheets for plotting a 2D line graph as shown in figures 3 and 4. The centroids represent the average emission for a particular year for each country, thus, providing spatial analysis.All the three states having 15 individual files are processed

using the above same procedure and then the results are saved in the .arff file format. Lastly, the results for each of the three countries are integrated to find out the percentage split between the two clusters. The final model is built in the Weka KnowledgeFlow providing a graphical image of the whole data mining process. The model build with Weka KnowledgeFlow is the same for all of the 3 countries as shown in the figure 2. The cluster identification for the countries were recorded and examined whether they had changed over time, thus, providing temporal analysis.

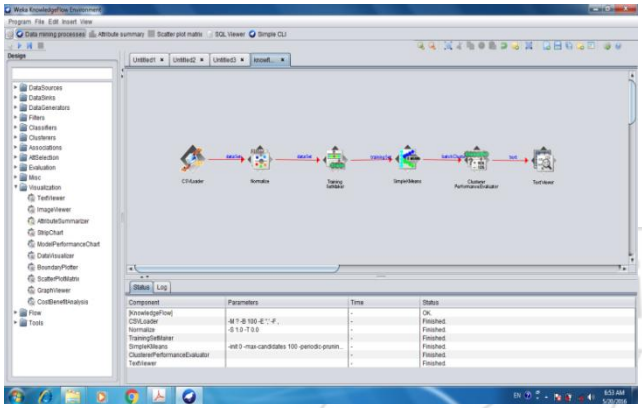


Figure 2: Weka KnowledgeFlow Model for the data mining process. It is the same for all of the 3 countries India, China and Indonesia.

4. Evaluation And Result

From the results obtained by following the CRISP-DM methodology onto the Weka Explorer, a number of conclusions are drawn:

1) The percentage split between the two clusters for each of the countries is 14% for high emitting cluster 0 and 86% for low emitting sector 1. The result so obtained is without any bias towards any of the clusters and is very good.
2) The emissions produced by India, as shown in the red line in Figure 3 are much lesser than Indonesia for all of the 15 years and lesser thanChina for only 6 of the 15 years though China is more developed country than India.

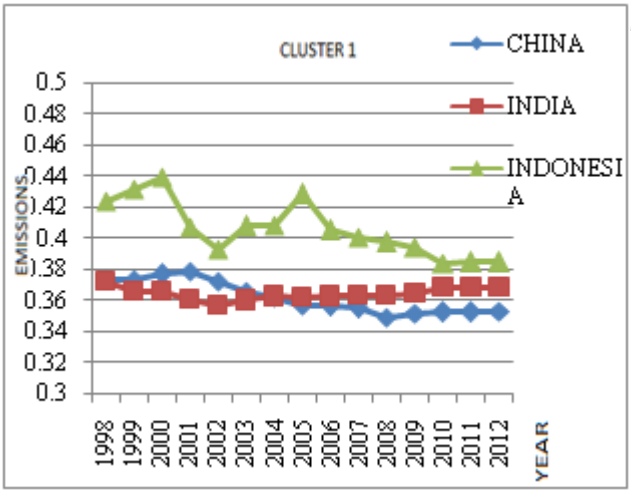


Figure 3: Showing the centroids of low emitting cluster 1

3) Also the point to be concerned about comes from the conclusion drawn from the high emitting cluster 0 where emissions for India are positive (that is, greater than 0).

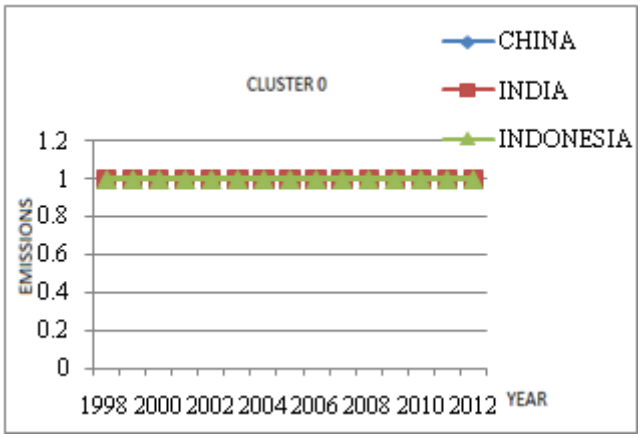


Figure 4: Showing the centroids of high emitting cluster 0

5. Conclusion

This research paper is one among the contributions made to the small amount of data mining applications in the environment sector. It also opens scope to discover and workout on more sectors contributing to Greenhouse Gas Emissions. The paper can act as an environment-related document for governmental as well as various environmental authorities working towards protecting the nation India from environmental catastrophes resulting into the social and economic losses so causedby analyzing the greenhouse gas emission levels as well as by knowing their position and status among some of the countries with almost similar population.

References

[1] Mariyam Kidwai and Garima Srivastava, *"Spatio-Temporal Clustering for Environmental Data: A Review"*, International Journal of Scientific Engineering and Research (IJSER), pg. 42-45, Vol. 4 Issue 4, 2016
[2] http://www.worldbank.org/en/news/feature/2013/06/19/india-climate-change-impacts
[3] http://www.climateemergencyinstitute.com/uploads/GLOBAL_WARMING_AND_ITS_IMPACTS_ON_CLIMATE_OF_INDIA.pdf
[4] A Report of the Intergovernmental Panel on Climate Change (IPCC), m.rediff.com/money/2007/jun/05clim.htm, *"Climate Change and its Impact on India"*, June 05, 2007, 10:29 IST.
[5] Alfredo Cuzzocrea, Mohamed MedhatGaber and Staci Lattimer, *"Spatio-Temporal Analysis of Greenhouse Gas Data via Clustering Techniques"*, Proceedings of the 2015 IEEE 19thInternational Conference on Computer Supported Cooperative Work in Design (CSCWD).
[6] http://www.sv-europe.com/crisp-dm-methodology/
[7] www.wikipedia.org