International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878, Impact Factor (2015): 3.791

Spatio-Temporal Clustering for Environmental Data: A Review

Mariyam Kidwai[#1], Garima Srivastava[*2]

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

Abstract: ***With the help of already known data, facts and figures, the process of extracting new useful information which was earlier not known is termed as data mining. The practice of data mining has been successfully implemented to a number of real-world applications. Researchers and data miners have done a lot of work on the environmental data and have generally done the spatio-temporal clustering of the data to extract some new and meaningful information regarding the environment related data which can be used to analyze environmental data by governmental authorities when decisions on such data are to be made in order to be provide deep insight of the data. [5] This paper aims to review different design methodologies, the protocols and government regulatory acts, bills and standards, spatio-temporal clustering techniques and the visualization tools used for the implementation of the environmental data.***

Keywords: Environmental Data Mining, Standards and Protocols, Design Methodologies, Clustering Methods, Spatio-temporal clustering, Visualization Tools

1. Introduction

Data mining is the mathematical approach of inspecting 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. This is leading to broaden the scope of data mining but at the same time challenging to the complexity and computation work in discovering new knowledge. Greenhouse gases namely carbon dioxide, hydro- fluorocarbons, methane, nitrous oxide, Ozone, water vapour etc. were always there in the atmosphere but from the last century, since their emissions and concentrations have gone through the noticeable increase, when they are released into the atmosphere, they track the infrared emissions of long wavelength from the sun leading to increased atmospheric temperature, thus causing the Greenhouse Gas Effect.

Greenhouse gas effect is a gradual phenomenon where there is an increase in the average temperature of the earth's atmosphere. This effect has caused the Global Environmental Change. Global warming has devastating impact on agriculture and crop yields, ecosystems, plant and animal habitats, coastal cities, human health, climate, global temperature etc. leading to the food insecurities, higher incidences of infectious diseases like dengue and malaria, frequent incidences of storms, floods and droughts. The greenhouse effect generally describes the condition where the excessive heat is trapped due to the increased level of carbon- dioxide concentration in the atmosphere. The carbon-dioxide absorbs a good amount of infrared radiations from the sun and simultaneously also does not allow it to escape into the atmosphere.

In a research, the greenhouse gas emissions for various country regions in terms of the major pollutants give the high and low emission zones.[5] In another research, air pollution, whose chemical composition is particulate matter is considered to be a phenomenon that could cause serious commercial problems in countries where economy mostly depends on tourism and agriculture.[7] High congestion of air pollution decreases crop yields and there are many crops that are profoundly sensitive to air pollution, thus, affecting supply and demand chain which means it affects both producers and consumers of food products.[7] Water Pollution resulting from the industrial or domestic activities, is a major problem in many countries, as approximately 25 million person die each year as a result of water pollution.[8][9] Ozone has been the main air-quality concern in HGB(Houston-Galveston-Brazoria) area for years.[11] Regional meteorological conditions combined with the variety of emissions from industry and transportation make the city a prime media for ground level ozone formation. [11][12] The aim of this paper is to provide a general insight of the methods, techniques and implementation tools that have been used by the researchers in mining the varying data of the environment.

2. Environment Protocols

The **Kyoto Protocol for Air Pollution** is a worldwide agreement that has been proposed with the objective of improving the worsening climatic factors specifically to cope up with the harmful effects of global warming by giving instructions to the countries to reduce the greenhouse gas emissions by maintaining the specified standards. The principle of the protocol is based on the presumption that there prevails global warming and that carbon dioxide emission is the major cause and that these emissions are human generated. It aims to reduce the raised concentrations of gases in the atmosphere to an extent that would prevent harmful degradation of the environment. [17]

The **Green House Gas (GHG) protocol** has been framed by the co-operations of World Resources Institute and World Business Council on Sustainable Development to tackle, measure, manage and control the climate change due to the greenhouse gas emissions across the world. The

standards are set as to cope up with the huge amount of greenhouse gas emissions due to the enormous industrial sectors in the world. The protocol is used by a number of governmental and public sectors, industrial and environmental groups, private and non- private organizations of the various countries. The GHG protocol is the foundation for all the environmental programs and GHG protocols and standards. Also, after the Kyoto Protocol, it's one of the most appreciable steps towards environmental change management and development. The spatio-temporal clustering of greenhouse gas emissions would be useful for such protocols for analyzing the current global situation.

The **Protocol on Water and Health** to the 1992 Convention on the Protection and Use of Trans-boundary Watercourses and International Lakes is the first major international legal approach for the prevention, control and reduction of water- related diseases in Europe. [18] So far, 36 countries have signed and 24 ratified it. [18] Signatories agreed to establish and maintain comprehensive national and/or local surveillance and early warning systems to prevent and respond to water- related diseases. [18] They also agreed to promote international cooperation to establish joint or coordinated systems for surveillance and early warning systems, contingency plans, and responses to outbreaks and incidents of water-related diseases and significant threats of such outbreaks. [18] WHO/Europe and the United Nations Economic Commission for Europe (UNECE) provide the joint secretariat for the Protocol, coordinating activities for its implementation. WHO handles the health aspects and UNECE the legal and procedural aspects. [18]

3. Design Methodologies

In the expanding field of data mining, there has been a call for a standard methodology or a simply list of best practices for the diversified and iterative process of data mining that users can apply to their data mining projects regardless of industry. [1] The SEMMA and CRISP-DM methodologies are generally used in the environmental data mining based papers since they have well documented and clearly defined steps along with the suitable flexibility.

1. **SEMMA** which is considered to be a general data mining methodology stands for Sample, Explore, Modify, Model and Assess.

Phases of SEMMA: The methodology consists of following five phases:

Sample: The sampling or selection of data is done in this process. The data so selected should be such as to be retrieved easily and efficiently used. Partitioning of data is also done in this phase.

Explore: Data is explored in this phase by discovering the relationships that exist among the variables in order to better understand the selected data.

Modify: In this phase, the data preparation takes place by modifying, formatting or transforming the data accordingly.

Model: In this phase, the modified data is modeled by applying various data mining techniques in order to obtain the desired results.

Assess: In this last phase, the reliability of the generated model is tested by implementations.

2. The Cross Industry Standard Process for Data Mining or **CRISP-DM**, founded by the European Strategic Program on Research in Information Technology is a robust and well- proven methodology. [1][2]

A structured approach is provided by the methodology for planning a data mining project. The analytics and data miners find it to be of powerful practicality, flexibility and usefulness while solving typical business issues. The CRISP-DM model is shown as follows:

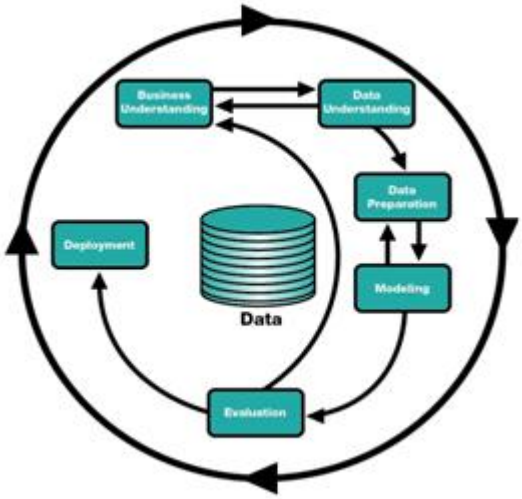

Figure 1: CRISP-DM Model [2]

Stages of CRISP-DM: The methodology consists of following six stages:

CRISP-DM stage one – business understanding: This very first stage focuses on understanding the business objectives and them converting them to fulfill data mining objectives. [3] CRISP-DM stage two – data understanding: The second stage includes collection of initial data by loading the necessary data for data understanding, describing, exploring, and verifying the data by examining the data quality. [4]

CRISP-DM stage three – data preparation: This phase consists of selecting appropriate data, cleaning it to gain quality output, constructing the data accordingly, and then integrating all the data obtained from different sources and finally formatting the data.

CRISP-DM stage four – modeling: Consisting of technique selection, test design, and building and assessing; the modeling phase is where the mining of the data takes place. The data is run through the chosen techniques, producing a model of the modified data. [5]

CRISP-DM stage five – evaluation: This phase ties in the business understanding which was established at the beginning, with the models of the data produced in the

modeling phase. It consists of evaluating the data mining results, reviewing the data mining process and determining next steps. [5]

CRISP-DM stage six – deployment: This phase is the application of the data mining results to business. [5] It includes plan deployment by determining strategy for deployment, plan monitoring and maintenance, producing final report by giving conclusions and reviewing project in terms of what went right and what went wrong, what was done well and what needs to be improved. [6]

4. Implementation Tools

Visualization tools are used by the researchers for the implementation of environmental data mining.

1. Java based 3D viewer: Java 3D is a scene graph (directed acyclic graph, DAG) based 3D application programming interface (API) for the Java platform.[10]
2. Google Earth : Google Earth is a virtual globe, map and geographical information program that was originally called Earth Viewer 3D created by Keyhole, Inc., a Central Intelligence Agency (CIA) funded company acquired by Google in 2004 (see In-Q- Tel).[13] It maps the Earth by the superimposition of images obtained from satellite imagery, aerial photography and geographic information system (GIS) onto a 3D globe.[13]
3. Waikato Environment for Knowledge Analysis (Weka): It is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand.[14] This original version was primarily designed as a tool for analysing data from agricultural domains,[15][16] but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research.[14] Weka supports several standard data mining tasks, more specifically, data pre-processing, clustering, classification, regression, visualization, and feature selection.[14] All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported).[5][14]
4. APCS-MLR: Absolute Principal Component Scores (APCS) and Multivariate Linear Regression (MLR): APCS has been used in many of pollution related Studies and Research such as air pollution [19][20],[21]-[24] and water quality assessment [25][26]. In the APCS-MLR, the APCS values produced by the rotated PCA are considered to be the independent variables and the measured concentrations of particular pollutant parameters are used as the dependent variables. The estimated contribution of each pollution source to the total concentration is determined by using the MLR.[9][25]

5. Clustering Technique

Spatio-temporal Clustering: Due to the advancements in geographical information systems and remote sensing networks, the availability and awareness of the dynamic and geographically distributed spatio-temporal data sets is increasing. This condition resulted in an increasing need for knowledge discovery in spatio-temporal data sets to be able to better cater to dynamic environment like moving objects (ex. Car), traffic management, forest fire, earthquake, hurricanes, air pollution, water pollution etc. These data sets might be collected at different locations at various instances or points of time in different formats. Therefore it has become very important to efficiently extract the spatio-temporal features, patterns and their relationships from such data sets to increase the significance of spatio-temporal data analysis and mining in many application domains, such as, weather forecasting, medical imaging, transportation monitoring and management, environment protection, crop sciences, geophysical activities and prediction, pollution incidents tracking, topological relationship patterns in agriculture and land-use management, traffic monitoring, planning afforestation, animal movements, supervising developments in embryology, changes in freezing level, earthquake histories etc. Spatio-temporal data mining is an emerging research area dedicated to the development and application of unconventional computational proficiencies for the study of large spatio-temporal databases. The practice of spatial-temporal data mining via clustering techniques has been successfully implemented to a huge number of real- world applications. This paper aims to review the spatio- temporal analysis and clustering of the greenhouse gas data. Spatio-temporal data mining is the practice of discovering hidden and implicit patterns or structures and geometry, the spatial (for example, distance, location, topology, direction, shape etc.) as well as the temporal (for example, valid timestamps, time intervals, before and after etc.) relations among them. The grouping of the resultant data sets is called spatio-temporal clustering. The k-means, k-medoid, spatio-temporal shared nearest neighbor (ST-SNN), ST-DBSCAN algorithms are used for the implementation work by several researchers. [7][9][11][12]

6. Conclusion

The spatio-temporal clustering techniques have been used to generate patterns of the environmental data such as air pollution data, water pollution data and greenhouse gas emissions data in various regions of the world in order to have a better understanding of the environmental processes and data. Such data is useful for governmental authorities for better planning and prevention techniques.

References

[1] https://en.wikipedia.org/wiki/SEMMA
[2] http://www.sv-europe.com/crisp-dm-methodology/
[3] http://www.sv-europe.com/business-understanding/
[4] http://www.sv-europe.com/data-understanding/
[5] Alfredo Cuzzocrea, Mohamed Medhat Gaber and Staci Lattimer, “Spatio -Temporal Analysis of Greenhouse Gas Data via Clustering Techniques”, Proceedings of the 2015 IEEE 19th International Conference on Computer Supported Cooperative Work in Design (CSCWD).

[6] http://www.sv-europe.com/deployment/
[7] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, Teh Ying Wah, Hoda Soltanian and Tutut Herawan, "Spatial and Temporal Clustering of Air Pollution in Malaysia: A Review", International Conference on Agriculture, Environment and Biological Sciences (ICFAE'14) June 4-5, 2014 Antalya (Turkey).
[8] B. Pimpunchat, W. L. Sweatman, G. C. Wake, W. Triampo, and A. Parshotam, "A Mathema tical model for pollution in a river and its remediation by aeration", Appl. Math. Lett, vol 22, pp.304-308, 2009.
[9] Jalal Karami, Abbas AliMohammadi, and Soroush Modabberi, "Analysis of the spatio -temporal patterns of water pollution and source distribution using the MODIS sensor products and multivariate statistical techniques", IEEE journal of selected topics in applied earth observations and remote sensing, vol. 5, no. 4, August 2012.
[10] https://en.wikipedia.org/wiki/Java_3D
[11] Sujing Wang, Tianxing Cai, Christoph F. Eick, "New Spatiotemporal Clustering Algorithms and their Applications to Ozone Pollution", 2013 IEEE 13 th International Conference on Data Mining Workshops.
[12] Texas Commission on Environmental Quality (TCEQ), Hourly ozone concentration data. http://www.tceq.state.tx.us. Accessed March 2010.
[13] https://en.wikipedia.org/wiki/Google_Earth
[14] https://en.wikipedia.org/wiki/Weka_(machine_learning)
[15] G. Holmes; A. Donkin; I.H. Witten (1994). "Weka: A machine learning workbench" (PDF). Proc. Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia. Retrieved 2007-06-25
[16] S.R. Garner; S.J. Cunningham; G. Holmes; C.G. Nevill- Manning; I.H. Witten (1995). "Applying a machine learning workbench: Experience with agricultural databases" (PDF). Proc. Machine Learning in Practice Workshop, Machine Learning Conference, Tahoe City, CA, USA. pp. 14–21. Retrieved 2007-06-25.
[17] Kyoto Protocol to the United Nations Framework Convention on Climate Change (http://unfccc.int/resource/docs/convkp/kpeng.pdf)
[18] http://www.euro.who.int/en/health-topics/environment- and-health/water-and-sanitation/protocol-on-water-and- health.
[19] H. Guo, T. Wang and P.K.K Louie, "Source apportionment of ambient non-Methane hydrocarbons in Hong Kong: Application of a principal component analysis/absolute principal component scores (PCA-APCS) receptor model," Environ. Pollution, vol. 129, pp. 489-498, 2004.
[20] G.D. Thurston and J.D. Spengler, "A quantitative assessment of source contributions to inhalable particulate matter pollution in Metropolitan Boston," Atmospheric Environment (1967), vol. 19, pp. 9-25, 1985.
[21] Y. Song, S. Xie, Y. Zhang, L. Zeng, L.G. Salmon and M. Zheng, "Source apportionment of PM2.5 in Beijing using principal component analysis/absolute principal component scores and UNMIX," Sci. Total Environ., vol. 372, pp. 278-286, 2006.
[22] M. Amodio, E. Andriani, I. Cafagna, M. Caselli, B.E. Daresta, G. deGennaro, A. Di Gillio, C.M. Placentino, and M. Tutino, "A Statistical investigation about sources of PM in South Italy," Atmos. Res., vol. 98, pp 207-218, 2010.
[23] T.W. Chan and M. Mozurkewich ,"Application of absolute principal component analysis to size distribution data: Identification of particle origins," Atmos. Chem. Phys., volume 7, pp. 887-897 2007.
[24] J. Miranda, I. Crespo and M. A. Morales, "Absolute Principal Component Analysis of atmospheric aerosols in Mexico city," Atmos. Aerosols, volume 7, pp. 14-18, 2000.
[25] F. Zhou, G.H. Huang, H. Guo, W. Zhang, and Z. Hao , "Spatio-temporal patterns and source apportionment of coastal water pollution in Eastern Hong Kong," Water Res., vol. 41 pp. 3429-3439, 2007.
[26] S. Su, J. Zhi, L. Lou, F. Huang, X. Chen, and J. Wu, "Spatio-temporal patterns and source apportionment of pollution in qiantang river(China) using neural-based modelling and multivariate statistical techniques," Phys. Chemistry Earth A/B/C, to be published