Air Pollution Monitoring using ANOVA and Big Data

D. Vishnu Naga Praveen[1], CH R Phani Kumar[2], K. Abhishek Anand[3], K. Chetana[4] and C. Raghava Reddy[5]

[1,2,3,4,5]Dept of ECE, GITAM Institute of Technology
E-mail: [1]vishnunagapraveen@gmail.com, [2]phanikrch@gitam.edu, [3]abhishekanand.kottur@gmail.com, [4]chetanakamjula9@gmail.com, [5]raghavac.reddy7@gmail.com

Abstract—
The project entitled "AIR POLLUTION MONITORING USING ANOVA AND BIG DATA" focuses on implementation of air pollution monitoring system. Now-a-days pollution monitoring is a challenging issue for many developing countries and the system should be cost effective as these are to be employed at many places. Knowing how much pollution is present in the air helps air quality agencies know what to do to protect public health. Some of the existing instruments for air pollution monitoring are Fourier transform infrared (FITR) instruments, gas chromatographs and mass spectrometers. However high cost, large size & maintenance cost made them unfavorable for monitoring applications on large scale.
In this project, first each sensor was tested after survey about market trends of a variety of sensors for detecting air pollution. The air pollution is monitored by interfacing the software program with the hardware module by getting the data of the concerned areas.
Data mining is a process of analyzing data from different perspectives and summarizing it into useful information. It allows users to analyze huge data from many different dimensions or angles. Technically, finding correlations or patterns among dozens of fields in large relational databases. DHT-11(digital humidity & Temperature) , MQ-9(co gas sensor) are used to collect readings of different climatic parameters using Aurdino UNO. The Aurdino is interfaced with the Bluetooth module for transmission and reception of data. Orange Canvas is the data mining tool used to analyze the data collected from the above sensors. The data is converted into a more meaningful graphical form using different widgets in the orange canvas software. ANOVA (Analysis of variance) is a statistical analysis tool that separates the total variability found within a data set. ANOVA test results can be used in a F-test on the significance of the regression formula overall.

Keywords*: Aurdino, Bluetooth HC-05, DHT, MQ-9, ANOVA, Data-mining.*

1. INTRODUCTION

The data is obtained from different spots of a geographical location. With the interfacing of the arduino and different kinds of sensors and Bluetooth module the data is logged. Data is analyzed using orange canvas data mining software. When we have only two samples we can use the t-test to compare the means of the samples but it might be unreliable in case of more than two samples. In this case ANOVA is more favorable. ANOVAs showcases the importance of factors by comparing the response variable means at the different levels. The main point of this research is showcasing the variations of different parameters further determining the F value.

2. METHODOLOGY

2.1 Collecting data from the different sensors and interfacing it with arduino and Bluetooth module–

We take readings from several sensors from several locations to obtain a wide variety of parameters on which we can perform ANOVA. Different sensors used are the DHT sensor(Digital Humidity & Temperature) and MQ-9 (CO gas sensor). We interface the sensors with the arduino as shown below.

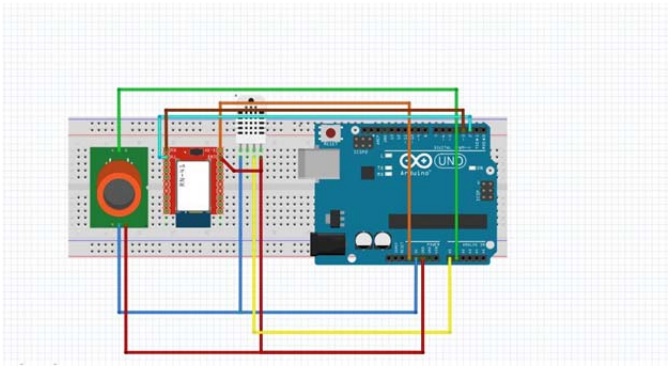

Fig. 1: Circuit Diagram of interfacing with the sensors

A Bluetooth module is interfaced to the arduino using the receiver and transmitter pins in the arduino. B-term is an application that displays the data obtained from the sensor arduino interface.

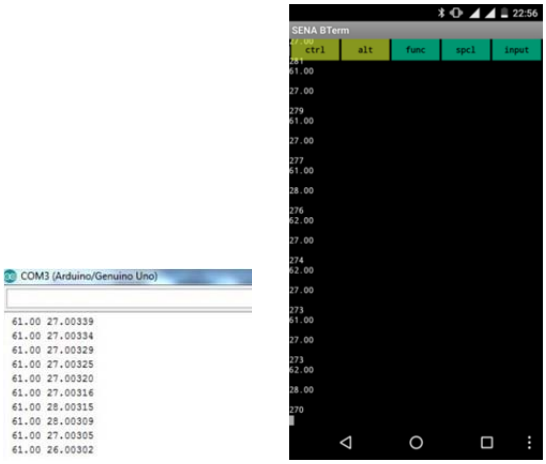

Fig. 2: Arduino serial port & Bluetooth interface display

2.2 Analysis using Data-mining –

The data mining tool used for this purpose is orange canvas. Using that data is analyzed and visualized.

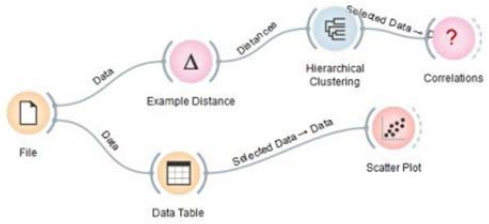

Fig. 3: Data mining linkage for Clustering & Correlation

The data is loaded into the file in the above figure and the following data is analyzed in following ways

2.2.1. Hierarchical Clustering –

The key technique of explaining data mining is clustering-separating the data into distinct groups based on some measure of similarity between two data instances through Euclidean (Pythagorean), Manhattan (sum of absolute difference between the coordinates).

Euclidean distance is calculated as:

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

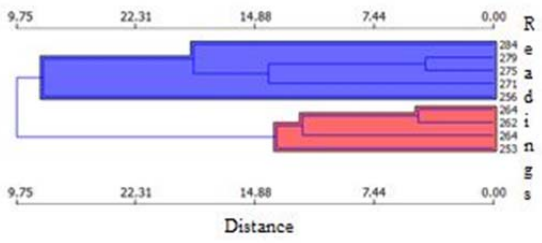

Fig. 4: Hierarchical Clustering Linkage

2.2.2 Correlation –

Correlation[2] is often used as a preliminary technique to discover relationships between variables. More precisely, the correlation is a measure of the linear relationship between two variables.

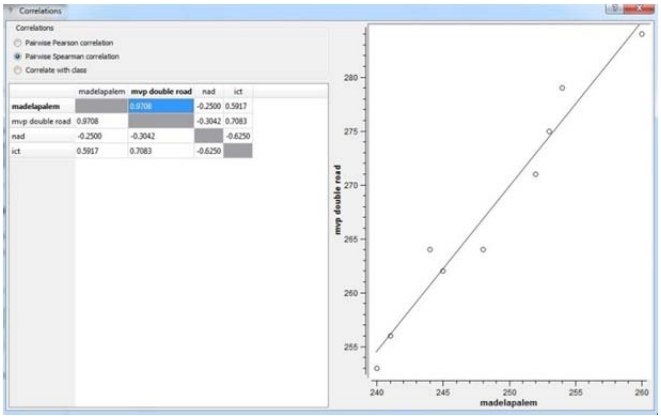

Fig. 5: Correlation using Spearman Correlation

2.3 Analysis using ANOVA –

The main aim in calculating ANOVA[1] is finding out the F-value of certain groups of parameters. In this case we consider 3 parameters of 4 different groups which indicates 4 different places as shown below

This is the live data collected at 4 different places denoted by l1,l2,l3,l4 at some peak hours of traffic. At first Parameter 1(humidity) variation is determined.

[5]Sum of Squares within Groups = total sum of $(X - \text{mean})^2$ of all 4 groups

Now, all the groups are combined together and the same procedure is done on the data to find out total sum of squares.[3][4]

Total Sum of Squares = **Sum of all groups together combined**

Sum of Squares between groups is calculated as = total sum of squares – sum of squares within groups

Degree of freedom = (number of samples in G1 - 1)+ (number of samples in G2 - 1)…..

$$F(\text{parameter } 1) = \frac{\text{Sum of squares between groups/degree of freedom}}{\text{sum of squares within groups/(observations} - \text{no.of groups)}}$$

L1	X-mean	(X-mean)^2	L2	X-mean	(X-mean)^2
60	-0.77778	0.60493	59	-0.11	0.01234
60	-0.77778	0.60493	60	0.889	0.79012
61	0.22222	0.04938	60	0.889	0.79012
61	0.22222	0.04938	59	-0.11	0.01234
61	0.22222	0.04938	59	-0.11	0.01234

61	0.22222	0.04938	59	-0.11	0.01234
61	0.22222	0.04938	59	-0.11	0.01234
61	0.22222	0.04938	59	-0.11	0.01234
61	0.22222	0.04938	58	-1.11	1.23456

L3	X-mean	(X-mean)^2	L4	X-mean	(X-mean)^2
59	-1.22222	1.4938	58	-1.77	3.1604
59	-1.22222	1.4938	59	-0.77	0.60493
59	-1.22222	1.4938	59	-0.77	0.60493
60	-0.22222	0.04938	60	0.222	0.04938
61	0.777778	0.60493	60	0.222	0.04938
61	0.777778	0.60493	60	0.222	0.04938
61	0.777778	0.60493	60	0.222	0.04938
61	0.777778	0.60493	61	0.222	1.49382
61	0.777778	0.60493	61	0.222	1.49382

F value of Parameter 1(humidity**)** = -0.123938967
In the same way for (parameter2) CO:

L1	X-mean	(X-mean)^2	L2	X-mean	(X-mean)^2
284	17	289	260	10	100
279	12	144	254	4	16
275	8	64	253	3	9
271	4	16	252	2	4
264	-3	9	248	-2	4
262	-5	25	245	-5	25
256	-11	121	244	-6	36
253	-14	196	241	-9	81
259	-8	64	253	3	9

L3	X-mean	(X-mean)^2	L4	X-mean	(X-mean)^2
277	1.55555	2.4197530	278	9	81
259	-16.4444	270.41975	277	8	64
260	-15.4444	238.53086	273	4	16
272	-3.44444	11.864197	260	-9	81
279	3.55556	12.641975	259	-10	100
293	17.55556	308.19753	262	-7	49
290	14.55556	211.86419	266	-3	9
280	4.555555	20.753086	264	-5	25
269	-6.44444	41.530864	282	13	169

F value[7] for parameter 2 (CO GAS)=1.088181101

In the same way for (parameter3)Temperature:

L1	X-mean	(X-mean)^2	L2	X-mean	(X-mean)^2
27	-0.33333	0.1111111	27	0.777	0.60493
27	-0.33333	0.1111111	26	-0.22	0.04938
27	-0.33333	0.1111111	26	-0.22	0.04938
27	-0.33333	0.1111111	26	-0.22	0.04938
27	-0.33333	0.1111111	26	-0.22	0.04938
28	0.666667	0.4444444	26	-0.22	0.04938
28	0.666667	0.4444444	26	-0.22	0.04938
28	0.666667	0.4444444	26	-0.22	0.04938
27	-0.33333	0.1111111	27	0.777	0.60493

L3	X-mean	(X-mean)^2	L4	X-mean	(X-mean)^2
28	-0.77777	0.6049382	28	1.444	2.08641
28	-0.77777	0.6049382	28	1.444	2.08641
29	0.222222	0.0493827	27	0.444	0.19753
29	0.222222	0.0493827	27	0.444	0.19753
29	0.222222	0.0493827	27	0.444	0.19753
29	0.222222	0.0493827	26	-0.55	0.30861
29	0.222222	0.0493827	25	-1.55	2.41975
29	0.222222	0.0493827	25	-1.55	2.41975
29	0.222222	0.0493827	26	-0.55	0.30864

F value for parameter 3 (TEMPERATURE) = 2.275362813.

3. CONCLUSION

We have used both practical approach and theoretical approach for the analysis of the Climatic variations due to pollution. We have gathered the data at four different locations at some peaks hours of traffic and have shown the pattern or trend of different parameters using the data mining technique and further for the theoretical approach we used the ANOVA analysis of variance in order to determine the F-value. F value determines the variation among group means. If the null hypothesis is satisfied, F have a value close to 1.0. Large F-value indicates variation among group means is more.

4. FUTURE SCOPE OF WORK

We have considered 9-10 readings at a particular place, likewise 4 different places are considered. In the same way if we can continuously monitor the level of variations of different parameters such as temperature, humidity, CO gas level we can control the pollution levels at certain places which shows a high level of variation compared to the normal readings.

REFERENCES

[1] Darrel L. Williams; James R. Irins; Brian L. Markham; Ross F. Nelson; David L. Toll; Richard S. Latty; Mark L. Stauffer,“ A Statistical Evaluation of the Advantages of LANDSAT Thematic Mapper Data in Comparison to Multispectral Scanner Data” DOI: 10.1109/TGRS.1984.350624

[2] W.M. Elliott, S. Kukunaris, J. Rohed, J. Liu, “Multifactor analysis of variance (ANOVA) investigation into GTEM to open area test site correlation for FCC Part 15 radiated emissions tests”, Electromagnetic Compatibility, 1998. 1998 IEEE International Symposium on (Volume:2)

[3] Reich, M.T.; Nelson, R.M.; Bauer-Reich, C. "The effect of EUT position on Gigahertz Transverse Electromagnetic (GTEM) cell correlation algorithms", Electromagnetic Compatibility, 2008. EMC 2008. IEEE International Symposium on, On page(s): 1 – 7

[4] Leslie Chandrakantha , “Learning ANOVA concepts using simulation”, American Society for Engineering Education (ASEE Zone 1), DOI: 10.1109/ASEEZone1.2014.6820644

[5] T. R. Black, "Simulation on spreadsheets for complex concepts: Teaching statistical power as an example," International Journal of Mathematical Education in Science and Technology, vol. 30, no. 4, pp. 473-481, 1999.

[6] Ch. R. Phani Kumar, B.UdayKumar, V. Malleswara Rao , Dsvgk Kaladhar. “Prediction of Effective Mobile Wireless Network Data Profiling Using Data Mining Approaches ” p-ISSN: 2324-9978 e-ISSN:2324-996X 2013; 2(1): 18-23.