ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

Air Pollution Clustering Using K-Means Algorithm in Smart City

Doreswamy[1], Osama A.Ghoneim[2], B R Manjaunath[3]

Department of Computer Science, Mangalore University, Mangalore, India[1]

Department of Computer Science, Mangalore University, Mangalore, India[2]

Department of Marine Geology, Mangalore University, Mangalore, India [3]

ABSTRACT: The vast amount of data produced by the Internet of Things (IoT) are considered of high business value, and data mining methods can be used to discover hidden valuable information from IoT data. Smart City is one of the most important applications of IoT. Smart city is currently dealing with rapidly increasing air pollution that result from variety of sources. The main cause of pollution is fumes gas from traffic system with a huge number of private vehicles. In order to help the city's environment people to deal with air pollution in smart ways, to find the best healthy area in the smart city which are suitable for living. In this paper we apply the K-means clustering algorithm on air pollution data level of pollution relying on available datasets generated from The CityPulse project [1]. The volume of data collected from each region of The CityPulse project can be extremely enormous and dynamic due to the number of mobile sensors deployed in the same location at the same time and their measurement frequency. This paper provides to viewer the actual level of pollution by position.

KEYWORDS: DATA MINING, CLUSTERING, K-MEANS, SMART CITIES AND AIR POLLUTION.

I. INTRODUCTION

Smart city data is big data. It is not only huge in volume, also it is multi modal, changes in format, representation form, quality and levels of dynamicity. City Pulse aims to offer large-scale real time solutions to interlink data from IoT and associated social networks and to achieve real-time information for the maintainable and smart city applications. The smart cities are evolving into larger ecosystems that were already disconnected. More and more services and applications in these projects are going to be online. Nowadays huge amounts of valuable data and sensor information still unused or restricted to certain service domains due to the large number of specific technologies and formats (like parking spaces, traffic information, bus timetables, waiting times at events, event calendars, environment sensors for pollution or weather warnings, GPS databases). The aggregation of information from various sources is typically done manually and the collective data is often static. CityPulse will speed up the creation and establishment of valid real-time smart city applications by accumulation two or more disciplines of knowledge-based computing and reliability testing.

Road traffic makes a significant provision to the following emissions of pollutants: benzene (C_6H_6), nitrogen dioxide (NO_2), carbon monoxide (CO), lead, Ozone (O_3), particulate matter (PM10 and PM2.5) and sulphur dioxide(SO_2). The impact of local air pollution on the environment and human health have been studied and well documented. We summarize the interaction and support chain of the people, traffic, air quality and health as Fig.1.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

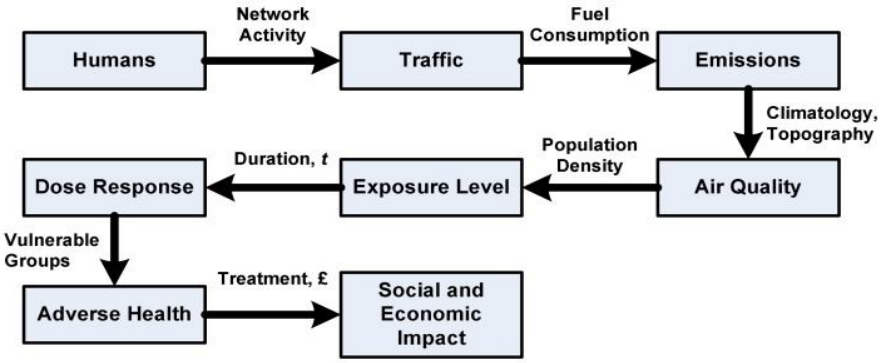

Fig.1: Effects of air pollution [2].

The figure shows that, increasing in car quantity and use in urban areas generates some chemical emissions to the air making the air pollution. With different climate situation, (effected by the wind, temperature, pressure, humidity, etc.), these pollutants pose different air qualities. When human beings expose to the fouled air (especially in the civil areas), driving in heavy traffic, nearby the freeways or at the 'downwind' areas, those people may suffer breathing problems and asthma attacks, which will lead to risk of heart attacks among people with heart disease. The effects of air pollution component on human health are listed below, Ozone (O_3), particulate matter (PM10 and PM2.5) and sulphur dioxide (SO_2).

- **Ozone (O_3)**

Scientific research shows that low-level ozone not only affects people either with impaired respiratory systems (such as asthmatics), or healthy adults and children as well. Exposure to ozone (O_3) for short duration (i.e. 6 to 7 hours), even at low levels, naturally reduces lung function and induces respiratory inflammation in normal. It can be accompanied by symptoms such as nausea, chest pain, coughing, and pulmonary congestion. Results from studies done on animal shows that frequent exposure to high concentration of ozone (O_3) for several months can lead to damage in the lung.

- **Carbon Monoxide (CO)**

Enters the bloodstream and reduces oxygen delivery to the body's organs and tissues. The health treatment from CO is most vital for those who suffer from cardiovascular disease. Also, healthy individuals are affected, but only at higher concentrations of exposure. Exposure to high CO concentration is related with reduced work capacity, visual impairment, and poor learning ability

- **Sulphur Dioxide (SO_2)**

The major health concerns associated with exposure to higher levels of SO_2 include effects on breathing, respiratory diseases, aggravation of existing cardiovascular disease, and alterations in pulmonary defences. Major subclasses of the people that are most sensitive to SO_2 include asthmatics and people who are suffering from cardiovascular disease or chronic lung disease (like emphysema or bronchitis).

- **Nitrogen Dioxide (NO_2)**

Nitrogen oxides are important in formation ozone and may affect both earthy and watery ecosystems. Nitrogen dioxide can aggravate the lungs and lower resistance to respiratory illnesses like influenza. The continued exposure to levels that is much higher than those normally found in the ambient air may cause raised rate of acute respiratory diseases in children.

- **Particulate Matter (PM-10 and PM-2.5)**

Main issues on people's health from the emissions of particulate matter are: Effects on breathing and respiratory systems, cancer, Lung disorders and early death. The elderly, children, and people, who are suffering from chronic lung disease, influenza, or asthma, tend to be specifically sensitive to the effects of particulate matter.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

II. RELATED WORK

Air pollutants, like Ozone (O_3), Nitrogen Dioxide (NO_2), Sulphur Dioxide (SO_2), Carbon Monoxide (CO), Particulate Matter (PM-10 and PM-2.5), have a definite impact on human living conditions. Nowadays, data mining algorithm such as cluster techniques are essentially used to analysis the impact of air pollution add to the relationship between weather conditions and air pollution. The most common clustering techniques in this area are the k-means, and hierarchical method.

A. K-means method:

Li, L. et al. [6] and Cervone, G. et al. [5] are studied the impact of air pollution by k-means. Also, Alex Mace et al. [7] presented modifying k-means clustering algorithm that can use both of the trajectory variables and the associated chemical value to classify source regions of definite chemical category.

B. Hierarchical method:

Seungmin Lee et al. [11] used an agglomerative hierarchical clustering method to study the source of and good meteorological conditions for high levels of PM10 in Seoul, Korea. Joseph H. Casola et al. [8] specified weather schemes via using distinct hierarchical clustering algorithms. Also, S. Yonemura et al. [9] and Charbel Afif et al. [10] reviewed the properties of pollutant gas levels by different hierarchical clustering methods.

III.METHODOLOGY

This analysis is based on the generated pollution data that has been collected from the “CityPulse project” and the URL is-“ http://iot.ee.surrey.ac.uk:8080/datasets.html”. This database consists of five air-pollution elements or attributes and they are Ozone (O_3), Carbon Monoxide (CO), particulate matter (PM), Sulphur Dioxide (SO_2) and Nitrogen Dioxide (NO_2). In addition to timestamp, longitude, and latitude. The detailed database format of one location in the CityPulse project is shown in the Table 1. Also, the location of CityPulse project corresponding to the longitude and latitude coordinates is shown in Fig.2.

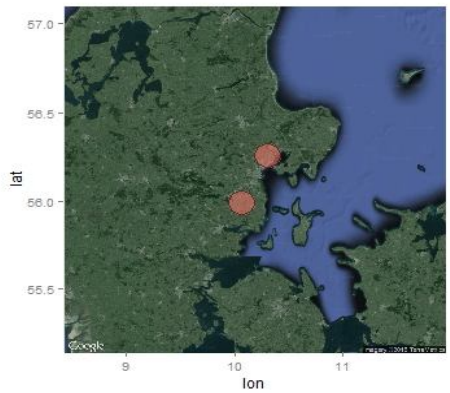

Fig.2. location of CityPulse project

Table 1: Original air-pollution Database from CityPulse project

.

Region	Ozone	Particullate_matter	Carbon_monoxide	Sulfure_dioxiode	Nitrogen_dioxide	longitude	latitude	timestamp
1	101	94	49	44	87	10.10499	56.23172	8/1/2014 12:05:00 AM
2	106	97	48	47	86	10.11659	56.22579	8/1/2014 12:10:00 AM
3	107	95	49	42	85	10.10711	56.21732	8/1/2014 12:15:00 AM
4	103	90	51	44	87	10.13978	56.21509	8/1/2014 12:20:00 AM

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

...	…	…	…	…	…	…	…	…
448	105	94	49	39	82	10.12501	56.23489	8/1/2014 1:10:00 AM
449	110	92	48	42	77	10.14507	56.21399	8/1/2014 1:15:00 AM

The CityPulse project contain 449 files each file for certain location in the city as shown in table 1 the longitude and latitude is fixed in each file. So in the beginning we aggregate these data using arithmetic mean for each attribute.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (1)$$

Where: $\bar{x}$ *is the average value of one of air pollution elements in cetain location.*
x_i *is the generated values of the air pollution element in the same location.*
n *is the number of the generated values for the air pollution element in the location.*

After that K-means clustering algorithm has been applied on the aggregated data, the best healthy area in the city (i.e. obtaining the smart environment in smart city) has been located.

A. K-means Algorithm

The k-means algorithm takes the input data set D and parameter k, and then divides a data set D of n objects into k groups. This partition depends upon the similarity measure so that the resulting intra cluster similarity is high but the inter cluster similarity is low. Cluster similarity is measured regarding the mean value of the objects in a cluster, which can be showed as the cluster's mean. The k-means procedure works as follows. First, it randomly chooses k of the objects, each of which initially defined as a cluster mean or center. For each of the remaining objects, an object is moved to the cluster to which it is the most similar, based on the similarity measure which is the distance between the item and the cluster average. It then calculates the new mean for each cluster. This process repeats until no change in the mean values in the clusters.

Algorithm: k-means.
Input: $E = \{e_1, e_2,..., e_n\}$ (set of objects to be clustered)
k (number of clusters)
Output: $C = \{c_1, c_2,..., c_k\}$ (set of cluster centroids)
$L=\{l(e) \mid e = 1,2,...,n\}$ (set of cluster labels of E)
Methods:

1. Randomly choose k points from the data set D as the initial cluster means (centroids);
2. Assign each object to the group to which is the most closest, based on the means values of the objects in the cluster;
3. Recalculate the mean value of the objects for each cluster;
4. Repeat the steps 2 and 3 until no change in the means values for the groups

IV.RESULTS AND DISCUSSION

The results of the K-means clustering algorithm on the air pollution data set from City Pulse project is given in the following tables.

Table 2: Mean distance between the clusters obtained using K-means algorithm, k=3

	Ozone	Particullate_matter	Carbon_monoxide	Sulfure_dioxiode	Nitrogen_dioxide
Cluster 1	102.6021	109.6021	111.4488	106.0446	98.23453
Cluster 2	110.3127	108.2743	106.2285	118.4539	98.23453
Cluster 3	125.0294	119.0904	121.0260	109.0264	109.87587

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

Table 3: Mean distance between the clusters obtained using K-means algorithm, k=4

	Ozone	Particullate_matter	Carbon_monoxide	Sulfure_dioxiode	Nitrogen_dioxide
Cluster 1	107.0683	119.7749	126.1922	116.5061	113.1374
Cluster 2	107.0097	108.4390	110.3419	98.9338	95.9647
Cluster 3	107.2753	101.2631	103.5042	120.0081	120.0316
Cluster 4	129.0777	117.2788	107.9018	109.3905	114.2417

Table 4: Mean distance between the clusters obtained using K-means algorithm, k=5

	Ozone	Particullate_matter	Carbon_monoxide	Sulfure_dioxiode	Nitrogen_dioxide
Cluster 1	107.6029	116.3321	94.43307	118.5035	110.25111
Cluster 2	101.9641	106.8681	116.17231	101.8949	95.98697
Cluster 3	115.5252	96.3166	108.52551	121.1870	120.26323
Cluster 4	130.2283	115.5439	114.95851	101.1870	109.94284
Cluster 5	107.9197	121.3714	125.17374	115.5040	118.58913

Table 5: Mean distance between the clusters obtained using K-means algorithm, k=6

	Ozone	Particullate_matter	Carbon_monoxide	Sulfure_dioxiode	Nitrogen_dioxide
Cluster 1	111.14315	120.61466	129.62295	116.84562	107.38115
Cluster 2	123.78890	110.82681	96.42087	119.62800	105.70056
Cluster 3	109.64336	97.21873	114.46163	121.94427	120.73909
Cluster 4	97.44348	121.47044	102.90947	113.15077	117.28840
Cluster 5	109.70825	104.62898	113.51247	95.15744	95.62908
Cluster 6	129.27222	116.24584	113.43613	101.11025	121.96706

Table 6: Mean distance between the clusters obtained using K-means algorithm, k=7

	Ozone	Particullate_matter	Carbon_monoxide	Sulfure_dioxiode	Nitrogen_dioxide
Cluster 1	108.2243	96.88181	97.98413	115.75647	117.29119
Cluster 2	111.4886	105.15550	131.81786	117.18624	117.42746
Cluster 3	101.0795	112.47313	112.03680	98.28057	96.99192
Cluster 4	105.7694	120.68373	120.01848	126.16312	103.99562
Cluster 5	127.8966	115.42956	100.95095	120.21328	104.28514
Cluster 6	130.1049	113.71992	116.34194	97.38804	113.524107
Cluster 7	106.8414	125.58255	109.34819	112.54993	129.78114

Table 7: Mean distance between the clusters obtained using K-means algorithm, k=8

	Ozone	Particullate_matter	Carbon_monoxide	Sulfure_dioxiode	Nitrogen_dioxide
Cluster 1	110.5566	130.57325	105.75646	118.57849	124.35556
Cluster 2	110.4003	108.33157	91.63987	117.01396	106.92799
Cluster 3	113.3194	121.56627	133.43648	119.47830	106.29373
Cluster 4	102.7995	103.91413	116.67895	129.33267	112.06851
Cluster 5	99.8783	112.75496	117.92125	102.86489	120.90732
Cluster 6	126.4190	98.28279	106.18010	115.80565	123.47248
Cluster 7	104.1712	103.45248	114.54593	99.54526	90.36685
Cluster 8	129.7986	117.25678	115.00825	99.55983	108.01890

Table 8: Mean distance between the clusters obtained using K-means algorithm, k=9

	Ozone	Particullate_matter	Carbon_monoxide	Sulfure_dioxiode	Nitrogen_dioxide
Cluster 1	133.22659	113.66339	101.11091	118.22697	111.31631
Cluster 2	111.61486	107.76818	114.15162	130.37970	107.44450
Cluster 3	108.09522	97.27836	118.34401	111.60389	124.80962

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

Cluster 4	114.82053	121.33967	136.22170	118.51311	109.39541
Cluster 5	90.16876	118.26718	117.59743	109.23675	99.42135
Cluster 6	114.03181	102.49507	111.83244	95.14472	93.75102
Cluster 7	106.19364	104.60168	91.71438	113.75989	112.22118
Cluster 8	127.73639	119.271181	118.02892	96.90242	113.09915
Cluster 9	106.41736	127.41143	108.29035	114.33172	129.01190

Table 9: Mean distance between the clusters obtained using K-means algorithm, k=10

	Ozone	Particullate_matter	Carbon_monoxide	Sulfure_dioxiode	Nitrogen dioxide
Cluster 1	108.50388	130.46822	111.04294	113.57209	131.13068
Cluster 2	114.71191	102.95795	112.14154	95.15947	93.84657
Cluster 3	129.12724	118.28160	117.18714	97.55066	113.80135
Cluster 4	115.90357	111.63613	138.91796	122.28279	112.76867
Cluster 5	100.02959	107.48233	92.11272	116.41592	114.40066
Cluster 6	102.93000	102.95874	119.27416	105.36514	122.55327
Cluster 7	89.66607	118.30523	118.43308	105.96982	99.67516
Cluster 8	119.45119	97.08002	109.04265	124.00820	121.91692
Cluster 9	127.47206	114.86291	97.06541	119.05108	102.96825
Cluster 10	109.20916	124.17216	118.14208	123.34767	102.08294

Ozone (O_3) is not emitted directly into the air by specific sources. Ozone (O_3) is created as a result of the effect of sunlight on nitrogen oxides (NOx) and volatile organic compound (VOC) emanations in the air. Often these "precursor" gases are emitted in certain location, but the real chemical reactions, encouraged by sunlight and temperature, take place in another. Mixed emissions from cars and stationary sources can be moved hundreds of miles from their origins, forming high ozone concentrations over very large regions. So we will focus in the analysis on the ozone (O3) average level in order to find the healthy and unhealthy area in the City Pulse project.

Using K-means algorithm grouped the CityPulse project into k groups. Each group has it's air pollution characteristics from these results the most healthy area for living was obtained at k=10 on cluster 7 where the Ozone (O_3) average is 89.66607ppb.This location is found at (longitude=10.17142 and latitude=56.15884). Also this area will have the less traffic if it is compared to other location in the city plus project. On the other hand, the most unhealthy and highly traffic in the city plus results is located at (longitude=10.18129 and latitude=56.16788) this area recorded the highest average level of Ozone (O_3) = 133.22659ppb.Fig.3.sohws the results of k-means at k=10

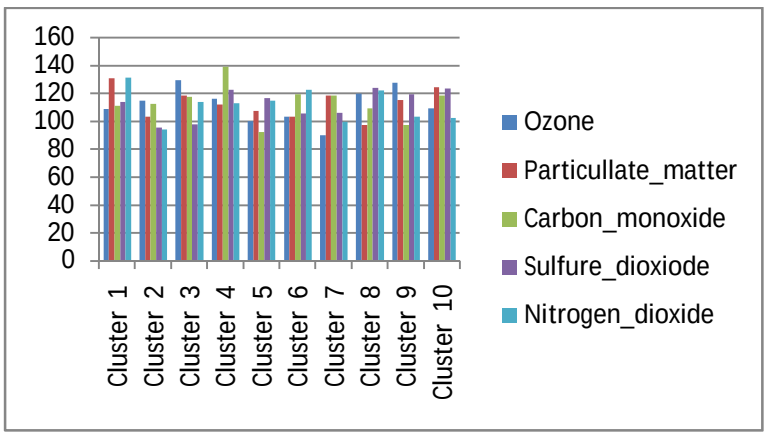

Fig .3.k-means results, k=10.

In Fig.3. it is notable that cluster 7 contain the minimum concentration level of Ozone (O_3), also the level of Nitrogen dioxide (NO_2) is less when it is compared to other clusters, which reflect the relation between the Ozone (O_3) and Nitrogen dioxide (NO_2).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

V.CONCLUSION

In this paper K-means clustering technique had been applied on the air pollution data set from City Pulse project. K value changed from 3 to 10. The cluster analysis is focused on the Ozone (O_3) average concentration. Healthy and unhealthy location in the City Pulse project has been determined which helps in getting smart environment in smart city.

In future, other data sets from the City pulse Project will be aggregated with the air pollution data in order to make city Pulse project smarter.

REFERENCES

1. "CityPulse project" and the URL is http://iot.ee.surrey.ac.uk:8080/datasets.html cited [7-09-2015].
2. Yajie Ma, Mark Richards, Moustafa Ghanem, Yike Guo 1 and John Hassard, "Air Pollution Monitoring and Mining Based on Sensor Grid in London" Sensors 2008, 8, 3601-3623; DOI: 10.3390/s8063601.
3. B. Ojeda-Magaña1, M. G. Cortina-Januchs2, J. M. Barrón-Adame, J. Quintanilla-Domínguez, W. Hernandez,A. Vega-Corona, R. Ruelas1 and D. Andina "Air pollution Analysis with a PFCM Clustering Algorithm Applied in a Real Database of Salamanca (Mexico)" 978-1-4244-5697-0/10/$25.00 ©2010 IEEE
4. Wei Tian1, Yuhui Zheng1, Runzhi Yang2, Sai Ji1 and Jin Wang1" A Survey on Clustering based Meteorological Data Mining" International Journal of Grid Distribution Computing Vol.7, No.6 (2014), pp.229-240
5. G. Cervone, P. Franzese, Y. Ezber, Z. Boybeyi, F. Bonchi, B. Berendt, F. Giannotti, D. Gunopulos, F. Turini, C. Zaniolo, N. Ramakrishnan and X. Wu, "Risk Assessment of Atmospheric Hazard Releases using K-means Clustering", Proceedings of IEEE International Conference on In Data Mining Workshops, (2008) December 15-19; Pisa, Italy.
6. L. Li and S. Cheng, "A Calculated Methodology of Regional Contributions Based on MM5-CAMx in Typical City: A 2006 Case Study of SO2 and Sulfate", Proceedings of the 4th International Conference on Bioinformatics and Biomedical Engineering, (2010) June 18-20; Chengdu, China.
7. A. Mace, R. Sommariva, Z. Fleming and W. Wang, in Adaptive K-means for clustering air mass trajectories, Edited H. Yin, W. Wang and V. Rayward-Smith, Springer Berlin, Heidelberg, vol. 6936, (2011), pp. 1-8.
8. J. H. Casola and J. M. Wallace, Journal of Applied Meteorology and Climatology, vol. 46, (2007).
9. S. Yonemura, S. Kawashima, H. Matsueda, Y. Sawa, S. Inoue and H. Tanimoto, Theoretical and Applied Climatology, vol. 92, no. 1, (2008).
10. C. Afif, A. L. Dutot, C. Jambert, M. Abboud, J. Adjizian-Gerard, W. Farah, P. E. Perros and T. Rizk, Air Quality, Atmosphere & Health, vol. 2, no. 2, (2009).
11. S. Lee, C. H. Ho and Y. S. Choi, Atmospheric Environment, vol. 45, no. 39, (2011).