j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 4, No:2 2021

Big Data Visualization by MapReduce for Discovering the Relationship Between Pollutant Gases

Y. A. Alsultanny

College of Engineering, Uruk University. Baghdad, Iraq

alsultanny@yahoo.com

Abstract Big data mining and pollution are extremely important issues in todays. An innovative method in this study was used for visually discovering the relationship between pollutant gases by MapReduce. One dimensional, two-dimensional, and three-dimensional visualization used to visualize the data, that was processed as an hourly reading for one year from an air quality monitoring station to study the behaviors of pollutant gases distribution, and to show graphically the distribution of one, two, or three gases. The number of readings used in this paper are 8760 hourly readings for each of the five pollutant gases under this study. Pearson correlation used to explore numerically the correlation between the pollutant gases, and eta factor used to evaluate the effect of one gas on the other pollutant gases. We found out by both methods, visually and numerically the same facts that related between the pollutant gasses. The ozone has a moderate negative correlation of value -0.622 with nitrogen dioxide, and weak negative correlation of value -0.248 with carbon monoxide, and -0.155 with carbon dioxide. Ozone has approximately no correlation of value .060 with silver dioxide. The carbon monoxide has moderate positive correlation of value 0.364 with carbon dioxide. The eta factor between ozone and nitrogen dioxide is very weak of values 0.292, and 0.009 with Sulphur dioxide, this proved an important fact that the ozone, nitrogen dioxide, and Sulphur dioxide sources are different. The study recommends that each country must analysis visually and numerical the big data that was collected yearly from the monitoring stations to control the pollution gases especially near the large industrial factories.

Crossref 10.36371/port.2021.2.3

Keywords: *Air quality; data mining; d gas pollution; carbon dioxide; correlation.*

1. INTRODUCTION

Big data mining is processing a large and complex sets of data, accumulated over time, the amount of data growing rapidly anywhere. Big data is a term used to describe some of current directions in technology, as a concept that must be taken into consideration in data analysis. It is important to note that most of the big data is unstructured data, and unorganized, that is difficult to fit as the usual databases [1,2]. In vast scientific spheres, state-of-the-art sensors that gather big data are always utilized. For instance, sensors are commonly used to obtain valuable physical, chemical, and biological data. Consequently, this offers improvement in the socio-economic well-being. Equally important in its applications for weather forecasting, monitoring, and timely responding to natural disasters and climate change .[3,5]

Big data is the information assets characterized by a high volume, velocity, and variety that required specific technology and analytical methods for its transformation into useful information [6]. Nature protecting and reducing environment pollution by processing data is one of the major achievements of big data analysis. Big data analysis can be an efficient tool that provides information for a sustainable economic and social future [7]. China for example started big data analysis for ecological and environmental protection.[8]

The demand for oil consumption was increasing, at the same times the air quality monitoring stations measured pollutant gases; these stations measure the pollutant gases such as ozone (O3), nitrogen dioxide (NO2), sulphur dioxide (SO2), carbon dioxide (CO2), carbon monoxide (CO), and others. The air quality stations usually are located near possible polluted areas such as oil refineries and factories. These stations also measure the meteorological parameters such as temperature, relative humidity, wind speed, wind direction, and many other parameters, that are important to monitor air quality. These stations working 24 hours, there is one reading every 5 minutes, these readings of the Arab Gulf region are available for this paper for more than two decades ago from many stations. The data that was used in this research study was pre-processed and analysed for only one air quality monitoring station in 2020.

LITERATURE REVIEW

Data mining is usually known as the technique to get useful knowledge out of databases. Data mining is the procedure that travels throughout data to discover unknown relations among

Alsultanny, Y. A. (2020) Big Data Visualization by MapReduce for Discovering the Relationship Between Pollutant Gases. *Journal Port Science Research*, *4*(2). pp 69–76. https://doi.org/10.36371/port.2021.2.3

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 4, No:2 2021

the data that are interesting to the user of the data [9], [10]. Data mining is used to describe the process of selecting, exploring, and modelling very large quantities of different types of data. This process is actually to figure out, if there are any original relationships or regularities that are still unclear with the aim of obtaining clear, obvious, relevant, and useful results for the data-miners [11,13].

Data mining includes techniques such as classification, clustering, and prediction. Classification is the process of finding a set of models that describe and distinguish data classes or concepts. Clustering is concerned with the problem of decomposing or partitioning a data set into groups so that the elements in one group are similar to each other and are as different as possible from the elements in other groups [14,15]. There are many algorithms for data mining such as C4.5, k-Means, MapReduce, Hadoop, SVM, A priori, EM, kNN, Naive Bayes algorithm, CART, Naïve Bayes, Random Forest, and Artificial Neural Networks [16,19].

The data mining algorithms can be used for big data analysis to improve air quality analysis [20]. MapReduce algorithm simplified data processing on large clusters. This algorithm allows splitting of a single computation task to multiple nodes for distribution processing [21]. MapReduce consists of two processes map and reduce, in the map process performs filtering and storing, and in reduce process performs a summary operation. In the map process, the master node takes the problem divides the problem into smaller sub-problems and distributes them to the worker nodes. The worker node takes care of processing smaller problems and carries the answer back to its master node [21]. Individual nodes perform the computing operation and return the results to the reduce function. The reduce function collects the individual results of the computation to generate a final output.[22]

MapReduce used by Shirwadkar in 2017 [23], for processing Texas air quality data. Data visualization is one of the significant components of data analytics in the age of big data. Visualization includes three presenting results one-dimensional (1D), two-dimensional (2D), and three-dimensional (3D) plots. The 3D trajectory reconstruction offers important opportunities to visualize the data [24]. Visual analytics enables organizations to take raw data and present it in a visual form that represents the data. Visualization of big data is bound to lead to some challenges and the opportunity for success with a data visualization strategy is much greater .[25]

Data visualization describes any effort to help people understand the significance of data by placing it in a visual context. It helped data engineers and scientists keep track of data sources and do basic exploratory analysis of data [26,27]. In large-scale data visualization, many researchers used feature extraction and geometric modelling to greatly reduce data size before actual data rendering [28]. Visualization is very important tool, and visualization can be used in mining big data, by showing the concentrations of gases [29]. One of the most important data visualization utilization is in the centralized control center, such as central control centre, the responsibility of the central control centre is to collect and monitor all the information in real time, to take an appropriate decision [30].

2. MATERIALS AND METHODS

Data pre-processing helps to improve the accuracy of the analysis, in this process, data may be deleted or edited to eliminate the redundant data to improve the data quality [31]. Therefore, pre-processing applied on the collected data. The data of this paper collected from one of the air quality monitoring stations in Arabian gulf region. The data are for the five pollutant gases O3, NO2, SO2, CO2, and CO. These data were pre-possessing to deal with missed data. There is one reading for each five minutes, the total readings for the five gases are 12*24*365*5=518400, the average of the readings for each one hour was calculated and represented as an hourly reading, these hourly readings used in the data analysis in this paper. Therefore, the number of readings used in the data analysis is: 24*365=8760 readings for each of the five gases in one year. Therefore, the MapReduce is the best method to show graphically the relationships between the five gases .

The descriptive method was used to visualize and understanding what has already happened in year 2020 by using MapReduce. Visualizing the data is very important to explore the nature of the data. Without visualizing it could not be easy to figure out the relation between the pollutant gases during the time series. Visualization is one of the important techniques that helps decision makers in identify out any increases in the concentrations of the pollutant gases, that could have bad effect on the climate over time in a very fast and undertakable way. The second approach used is the quantitative method, by using statistical analysis such as, Pearson correlation to measure the correlation between the pollutant gases, and eta factor to find out the effect of one pollutant gas on the other pollutant gases.

3. RESULTS AND DISCUSSION

Data visualization is an important tool to represents the data in graphical format. It enables decision makers to understand graphically the distributions, patterns, and trends of the readings, and make their decisions faster and accurate. Fig. 1 shows the five 1D charts, these charts are representing the 8760 readings for each gas during 12 months in 2020. The x-axis represents hours (00:00-23:00), the y-axis represents gas concentration. The chart of the O3 shows that its highest concentration distribution during the day's hours from noon to 8:00 pm. The NO2 have the lowest concentration distribution during the day hours especially from 10:00 am to 16:00pm. The chart of SO2 shows its highest concentration distribution is during 7:00 am to 14:00 pm and during 18:00 am to 22:00 pm. The chart of CO2 shows its concentration

Alsultanny, Y. A. (2020) Big Data Visualization by MapReduce for Discovering the Relationship Between Pollutant Gases. *Journal Port Science Research*, *4*(2), pp 69-76. https://doi.org/10.36371/port.2021.2.3

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 4, No:2 2021

distribution is approximately the same during the 24 hours. The chart of CO shows its highest concentration distribution is during 6:00 am to 10:00 am and during 19:00 am to 00:00 pm.

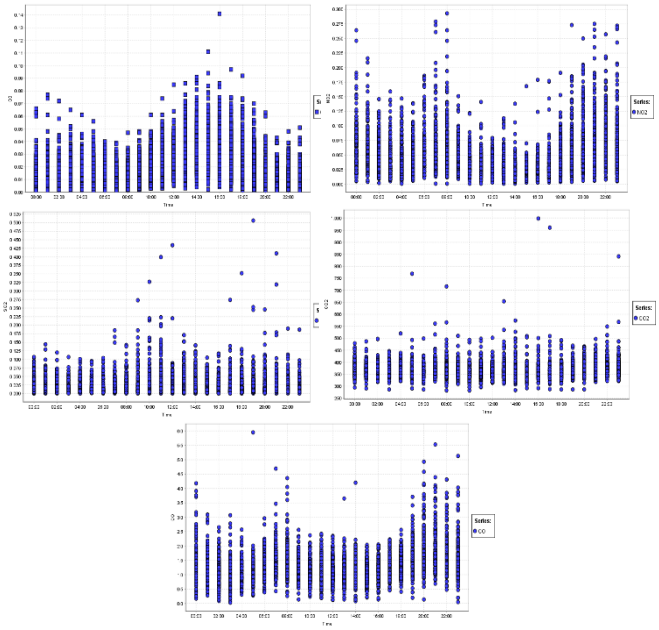
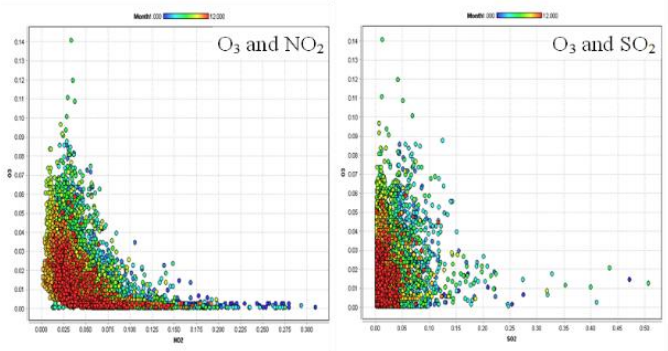


Fig. 1: *the daily 1D charts for the five pollutant gases in year 2020.*

Fig. 2 shows the ten 2D scatter charts to show the concentration distribution between the five gases, these charts are representing the 8760 reading for each gas during 12 months in 2020, the scale of the colours are for the months, January represented by the blue colour and December represented by the red colour. In these charts, the X-axis represents the independent variable that affects on the Y-axis that represent the dependent variable. The first row of charts in this figure shows the concentration distribution of NO_2, SO_2, CO_2, CO against O_3. The O_3 has an inverse relationship with NO_2, because the highest concentrations of both gases are distributed in reverse direction along the x-axis and y-axis. The other pollutant gases SO_2, CO_2, CO have a weak correlation with O_3, because they have no regular distribution. The second row of charts show NO_2 has no relation with SO_2, while has a positive relation with CO_2 and CO, this indicate the source of NO_2, CO_2 and CO_2 are the same. The third row of the charts show SO_2 has no relation with CO_2 and CO. The last chart in the fourth row shows that there is positive correlation between CO and CO_2.



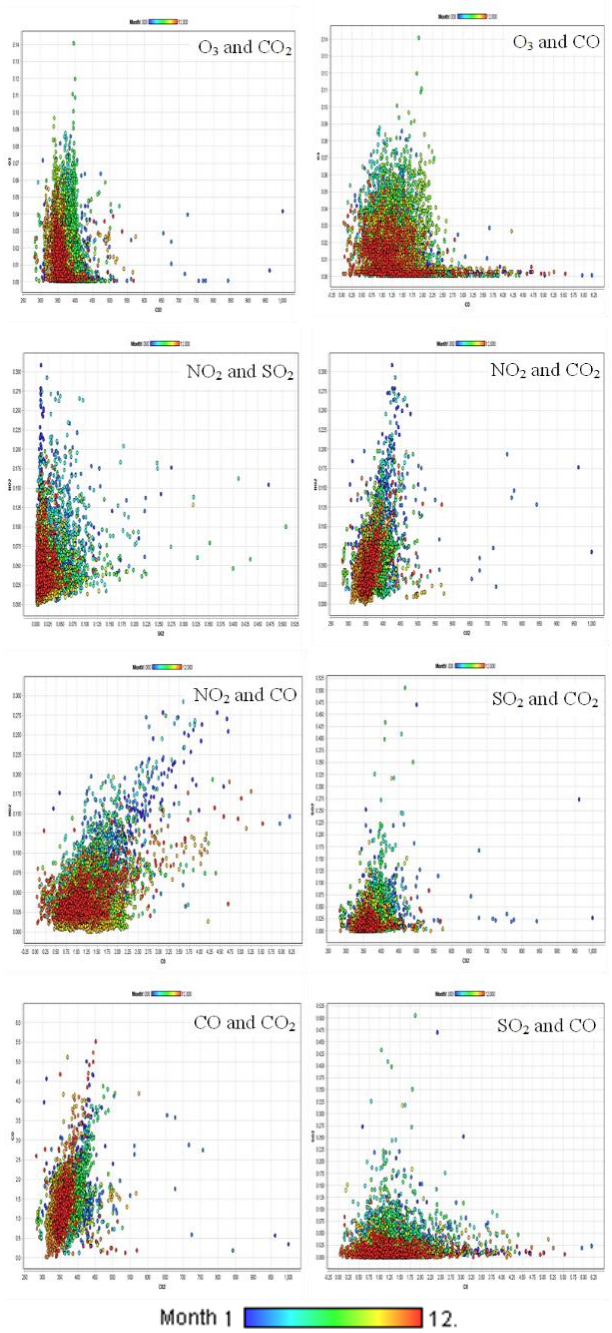


Fig. 2: *the 2D charts for the five pollutant gases in year 2020.*

Fig. 3 shows the ten 3D scatter charts, the three charts in the first row shows an inverse relation between O_3 and NO_2, while O_3 has no relation with SO_2. However, O_3 has positive relation with CO_2 and CO, because it has regular redistribution shape with them. The first two charts in the second row show O_3 has no relation with SO_2, but again has positive relation with CO_2 and CO, this also proved by the third chart in the second row. The three charts in the third-row show NO_2 has no relation with SO_2, but has a positive relation with CO_2 and CO, because it has regular redistribution shape with them. The last chart in the fourth row shows SO_2 has a positive relation with CO_2 and CO, because it has regular redistribution shape with them.

Alsultanny, Y. A. (2020) Big Data Visualization by MapReduce for Discovering the Relationship Between Pollutant Gases. *Journal Port Science Research*, *4*(2). pp 69–76. https://doi.org/10.36371/port.2021.2.3

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 4, No:2 2021

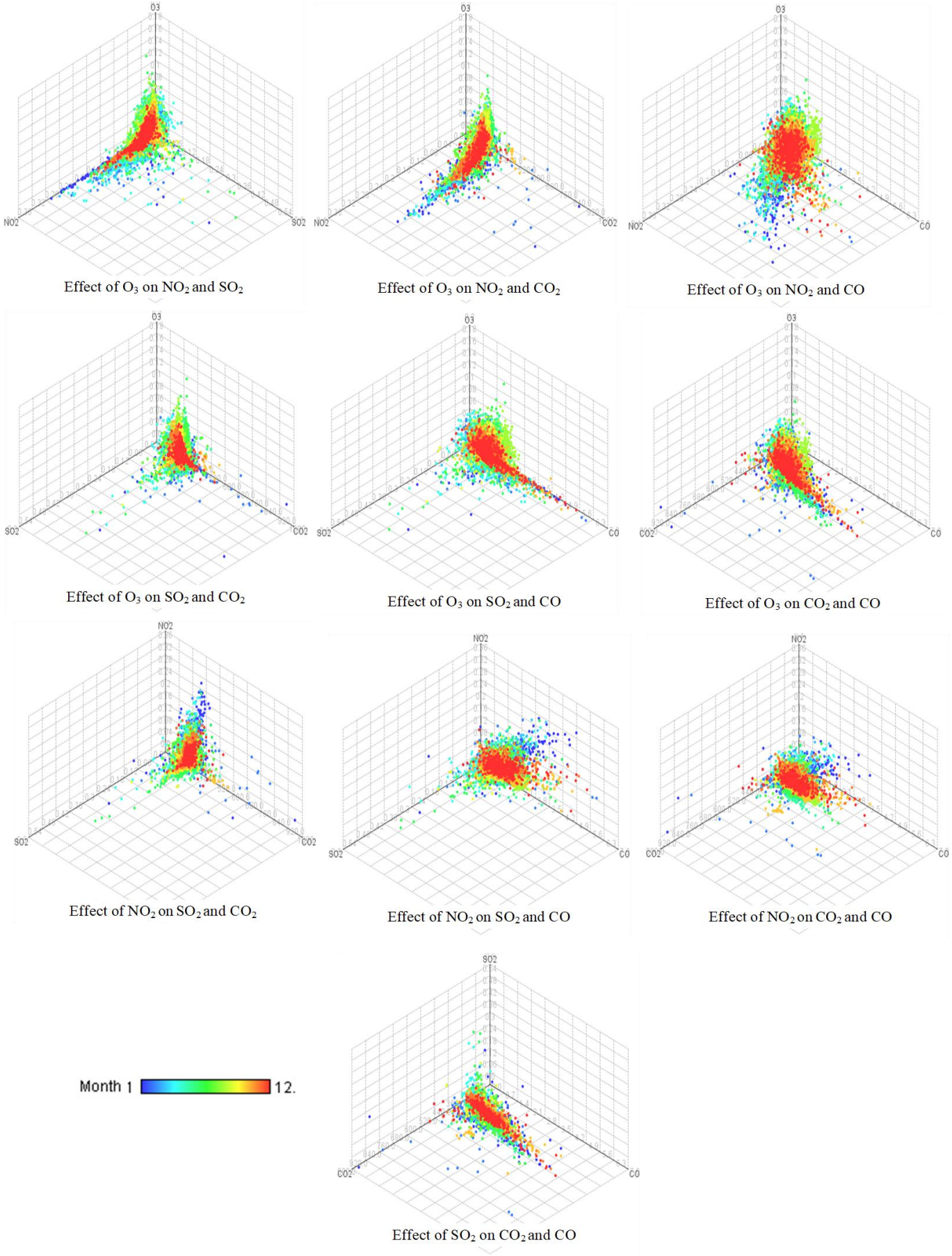


Fig. 3: *the 3D charts for the five pollutant gases in year 2020.*

Alsultanny, Y. A. (2020) Big Data Visualization by MapReduce for Discovering the Relationship Between Pollutant Gases. *Journal Port Science Research*, *4*(2), pp 69-76. https://doi.org/10.36371/port.2021.2.3

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 4, No:2 2021

To test the correlation between the five pollutant gases, the normality of data distribution test was applied by using one-sample Kolmogorov-Smirnov test. Table 1 shows that the five pollutant gases have non-normal distributions, because the Sig. (2-tailed) are 0.000 for the five gases. Therefore, the Spearman test was used to find out the correlation.

Table 1: *One-sample Kolmogorov-Smirnov test*

Type	*Function*	*O_3*	*NO_2*	*SO_2*	*CO_2*	*CO*
N		*7181*	*7181*	*7181*	*7181*	*7181*
Normal Parameters[a,,b]	*Mean*	*.01957*	*.04996*	*.02*	*370.45509*	*1.32626*
	S.D.	*.016240*	*.037060*	*.029*	*33.327782*	*.565714*
Most Extreme Differences	*Absolute*	*.126*	*.135*	*.228*	*.096*	*.091*
	Positive	*.094*	*.135*	*.206*	*.096*	*.091*
	Negative	*-.126*	*-.112*	*-.228*	*-.089*	*-.062*
Kolmogorov-Smirnov Z		*10.719*	*11.429*	*19.280*	*8.125*	*7.725*
Asymp. Sig. (2-tailed)		*.000*	*.000*	*.000*	*.000*	*.000*
a. Test distribution is Normal.						
b. Calculated from data.						

Table 2: *Correlations between the five gases*

Function	*Gas*	*Type*	*O_3*	*NO_2*	*SO_2*	*CO_2*	*CO*
Spearman's rho	*O_3*	*Correlation Coefficient*	*1.000*				
		Sig. (2-tailed)	*.*				
		N	*7181*				
	NO_2	*Correlation Coefficient*	*-.622***	*1.000*			
		Sig. (2-tailed)	*.000*	*.*			
		N	*7181*	*7181*			
	SO_2	*Correlation Coefficient*	*.060***	*.266***	*1.000*		
		Sig. (2-tailed)	*.000*	*.000*	*.*		
		N	*7181*	*7181*	*7181*		
	CO_2	*Correlation Coefficient*	*-.155***	*.422***	*.324***	*1.000*	
		Sig. (2-tailed)	*.000*	*.000*	*.000*	*.*	
		N	*7181*	*7181*	*7181*	*7181*	
	CO	*Correlation Coefficient*	*-.248***	*.333***	*-.009*	*.364***	*1.000*
		Sig. (2-tailed)	*.000*	*.000*	*.461*	*.000*	*.*
		N	*7181*	*7181*	*7181*	*7181*	*7181*
***Correlation is significant at the 0.01 level (2-tailed).*							

The correlation coefficient is represented by r in this paper. According to Vaske, Beaman, and Sponarski in 2017 [32] coefficient with a value r≥0.7 indicates a strong association between variables, moderate correlation when the values ±0.7>r≥0.3, and weak correlation for r <±0.3. Table 2 shows the correlation between the five pollutant gases. The O_3 clearly have moderate negative correlation -0.622 with NO_2, while it has weak correlations with SO_2, CO_2, and CO. The NO_2 has moderate positive correlation 0.422 with CO_2, and moderate positive correlation 0.333 with CO. The SO_2 has moderate positive correlation 0.324 with CO_2. The CO_2 has moderate positive correlation 0.364 with CO. All these results proved by the 2D and 3D charts.

To test the effect of one gas on the other, the paired samples T-test applied to measure the effect factor eta (η). Table 3 shows the effect of O_3 on NO_3 is 0.292 is very weak, this prove that the sources of these gases are different, also of O_3 have no effect on SO_2. While O_3 have high effect on CO_2 and CO, which indicates the pollution source that generate O_3 generate CO_2 and CO. The NO_2 has weak effect on SO_2 and high effect on CO_2 and CO, the SO_2 has high effect on CO_2 and CO this indicates the source of pollution for SO_2, CO_2 and CO is the same. The CO_2 has high effect on CO this indicates the source of pollution for them is the same.

Table 3: *Eta effect factor for the five gases*

N	*Gases*	*T value*	*df*	*η*
1.	*O_3 - NO_2*	*-54.366*	*7180*	*0.292*
2.	*O_3 - SO_2*	*-8.226*	*7180*	*0.009*
3.	*O_3 - CO_2*	*-941.839*	*7180*	*0.992*
4.	*O_3 - CO*	*-194.473*	*7180*	*0.840*
5.	*NO_2 - SO_2*	*54.986*	*7180*	*0.296*
6.	*NO_2 - CO_2*	*-942.224*	*7180*	*0.992*
7.	*NO_2 - CO*	*-197.65*	*7180*	*0.845*
8.	*SO_2 - CO_2*	*-942.104*	*7180*	*0.992*
9.	*SO_2 - CO*	*-195.064*	*7180*	*0.841*
10.	*CO_2 - CO*	*943.873*	*7180*	*0.992*

4. CONCLUSION

In this paper, three methods used to represent the concentrations, relationships between the five pollution gases, and effect of one gas on the other(s). The innovative method in this paper is the using MapReduce to show graphically the pollutant gases distributions of the hourly readings for one year. The 1D charts showed the distribution of each gas during the 24 hours, this type of charts helps in distributing the big data readings into groups have the same or similar values. The most important results of this type of charts, they proved that the O_3 highest level of concentration is mostly during the day hours from noon to 8:00 pm, While NO_2 have the reverse behaviour, its highest level of concentration starting from 8:00 pm till the 8:00 am of the next day. The distribution of CO is approximately similar to the NO_2 distribution, this means when the concentration of NO_2 increased the CO also increased, these two gases are very harmful to human life, and their main sources are burning the cars fuel, because most of the people in this country using their cars during night, due to the high temperature during the day hours, especially in summer that is staring from March till October. the CO_2 distribution is approximately the same during the 24 hours, because the main sources of this gas are the oil refinery stations and electricity power generation stations that are working 24 hours.

The 2D charts showed the combine distribution of each two different gases. The important results of this type of charts, they represent the distribution of each two gases during the 12 months in a year. The highest pollutant gases

Alsultanny, Y. A. (2020) Big Data Visualization by MapReduce for Discovering the Relationship Between Pollutant Gases. *Journal Port Science Research*, *4*(2). pp 69–76. https://doi.org/10.36371/port.2021.2.3

j. port. sci. res.
ISSN: 2616-7441 (Online)
ISSN: 2616-6232 (Print)
ISSN: 2616-7220 (USB)
INTERNATIONAL IDENTIFIER FOR SERIALS

Journal port Science Research
Available online www.jport.co
Volume 4, No:2 2021

concentrations were occurred during the summer months, they also showed there is divergence between O_3 and NO_2, this is another indicator that the sources of O_3 and NO_2 are different. Again, NO_2 and CO have directly direction, which is another prove that their sources are the same. The SO_2 don't have any regular behaviour with other four pollutant gasses.

The 3D charts showed the combine distribution of three different pollutant gases during the 12 months in a year. The important results of this type of charts showed that O_3 and NO_2, have relation with CO_2 and CO, this indicates the emission sources of O_3 and NO_2 also generate the pollutant gases CO_2 and CO during the periods of time, in spite of the sources of the O_3 and NO_2 are different.

The result of Spearman's correlation showed negative moderate correlation of value between O_3 and NO_2, but the eta effect factor between them is low, which is also prove that the sources of the emission of these two gases are not the same. The O_3 has high effect on CO_2 and CO, while it has weak correlation with them, because the sources of CO_2 and CO are the refinery stations and electricity power generation station, and another source is burning cars fuel, which also cause increase in the concentration of these two gases. The correlation between CO_2 and CO is moderate, but the eta effect factor is very high between them, this proves that the sources of emissions for these two gases are the same.

From all the above, the innovative methods of analysis used in this paper is very effective in monitoring the pollutant gases and measuring the correlation and determining the effect of one gas on the others. The MapReduce is one of the important methods to show graphically the distribution of big data, especially the air pollution data, because the size of these data collected by air quality monitoring stations are grow up, due to the importance of monitoring air quality each minute to reduce the effect of the pollutant gases on climate change.

ABBREVIATIONS USED IN THIS ARTICLE

η	*eta factor*
1D	*One-dimensional*
2D	*Two-dimensional*
3D	*Three-dimensional*
A priori	*A priori knowledge analytic*
am	*At morning*
C4.5	*Decision tree algorithm*
CART	*Classification And Regression Trees*
CO	*Carbon Monoxide*
CO_2	*Carbon Dioxide*
EM	*Expectation Maximization algorithm*
k-Means	*k- fixed number of clusters in a dataset*
kNN	*k-Nearest Neighbour*
N	*Number*
NO_2	*Nitrogen Dioxide*
O_3	*Ozone*
pm	*Past morning*
S.D.	*Standard Deviation*
Sig.	*Significant value*
SO_2	*Sulphur Dioxide*
SVM	*Support Vector Machine*

REFERENCES

[1] S. Lohr, (2012, Feb.). The Age of big data. New York Times. [Online]. Available: http://www.academia.edu/download/34393761/2_The_New_York_Times_on_The_Age_of_Big_Data.pdf

[2] V. Vadivu, "A Review on big data analytics," *IJSDR*, vol. 1, no. 10, pp. 264-266, Oct. 2016, [Online]. Available: http://www.ijsdr.org/viewpaperforall.php?paper=IJSDR1610044

[3] M. Zaki, M. Hartmann, N. Feldmann, and A. Neely, (2014). Big data for big business? a taxonomy of data-driven business models used by start-up firms. Cambridge Service Alliance, United Kingdom. [Online]. Available: https://cambridgeservicealliance.eng.cam.ac.uk/resources/Downloads/Monthly%20Papers/2014_March_DataDrivenBusinessModels.pdf

[4] R. Kune, K. Konugurthi, A. Agarwal, R. Chillarige, and R. Buyya, "The anatomy of big data computing," *Software Pract Exper*, vol. 46, no. 1, pp. 79-105, Oct. 2015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/spe.2374

[5] A. Honarvar and A. Sami, "Towards sustainable smart city by particulate matter prediction using urban big data, excluding expensive air pollution infrastructures," *Big Data Res.*, vol. 17, pp. 56-65. Sept. 2019, [Online]. Available: https://doi.org/10.1016/j.bdr.2018.05.006, https://www.sciencedirect.com/science/article/pii/S2214579617302587

[6] A. De Mauro, M. Greco, M. Grimaldi, and G. Nobili "Beyond data scientists: a review of big data skills and job families," *in Proc. of the 11th International Forum on Knowledge Asset Dynamics, IFKAD 2016.* June 15-17, 2016, pp.1844-1857, Dresden, Germany. [Online]. Available: https://scholar.google.com/citations?user=Wj1oWWEAAAAJ&hl=en

[7] D. Helbing, S. Frey, G. Gigerenzer, E. Hafen, M. Hagner, Y. Hofstetter, and V. Zicari, (2017, Feb.). Will democracy survive big data and artificial intelligence, Scientific American, a Division of Nature America Inc., USA. [Online]. Available: https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/

[8] B. Zhang, R. M. Hughes, W.S. Davis, and C. Cao, "Big data challenges in overcoming China's water and air pollution: relevant data and indicators," *SN Appl. Sci.* 3, 469, March 2021. [Online]. Available: https://doi.org/10.1007/s42452-021-04448-0

[9] Y. Alsultanny, Université de Bourgogne, France "Comparison between data mining algorithms implementation," in *Proceedings of the International Conference on Digital Information and Communication Technology and its Applications. DICTAP2011*, Université de Bourgogne, Dijon, France June 21-23, 2011, part II, CCIS 167, pp. 628-641. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-22027-2_52

[10] S. Kumar, and K. Kaur, (2016). Review of data mining (knowledge discovery) in the future. *IJARCS*, vol. 7, no. 6, 269-272, 2016. [Online]. Available: http://www.ijarcs.info/index.php/Ijarcs/article/view/2778

[11] P. Giudici, and S. Figini, "Applied data mining for business and industry," 2nd Edition, Wiley & Sons Inc., New Jersey, USA, 2009. [Online]. Available: https://www.wiley.com/en-us/Applied+Data+Mining+for+Business+and+Industry%2C+2nd+Edition-p-9780470058862

[12] R. Kimball, and M. Ross, "The data warehouse toolkit: the complete guide to dimensional modeling," 3rd Edition, Wiley & Sons, New Jersey, USA, 2013. [Online]. Available: https://www.amazon.com/Data-Warehouse-Toolkit-Complete-Dimensional/dp/0471200247

[13] G. Shmueli, C. Bruce, I. Yahav, R. Patel, and C. Lichtendahl, "Data mining for business analytics: concepts, techniques, and applications in R," John Wiley and Sons, New Jersey, USA. 2017. [Online]. Available: https://www.wiley.com/en-us/Data+Mining+for+Business+Analytics%3A+Concepts%2C+Techniques%2C+and+Applications+in+R-p-9781118879368

[14] H. Witten, E. Frank, A. Hall, and J. Pal, "Data mining: practical machine learning tools and techniques," 4th Edition, Elsevier, Amsterdam, Netherlands, 2016. [Online]. Available: https://www.elsevier.com/books/data-mining/witten/978-0-12-804291-5

[15] K. Purohit and K. Sharma, "Development of data mining driven software tool to forecast the customer requirement for quality function deployment," *IJBAN*, vol. 4, no. 1, pp. 56-86, 2017. [Online]. Available: https://pdfs.semanticscholar.org/3f65/7e0edefd127b54ce7af46de8a3a62f9672d5.pdf

[16] X. Wu, V. Kumar, R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z. Zhou, M. Steinbach, D. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowl Inf Syst.* 14(1): 1-37, 2008. [Online]. Available: https://link.springer.com/article/10.1007/s10115-007-0114-2

[17] J. Han, and M. Kamber, "Data mining concepts and techniques," 3rd Edition, Elsevier. Amsterdam, Netherlands, 2011. [Online]. Available: http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

[18] B. Srinivasan and P. Mekala, Mining social networking data for classification using REPTree. *IJARCSMS*, vol. 2, no. 10, pp. 155-160, 2014. [Online]. Available: https://www.academia.edu/9421415/Mining_Social_Networking_Data_for_Classification_Using_Reptree

[19] A. Nagar, "A Comparative study of data mining algorithms for decision tree approaches using WEKA tool," *AENSI*, vol. 11, no. 9, pp. 230-241, 2017. [Online]. Available: https://go.gale.com/ps/anonymous?id=GALE%7CA505467547&sid=googleScholar&v=2.1&it=r&linkaccess=abs&issn=19950772&p=AONE&sw=w

[20] E. Dragomir, M. Oprea, M. Popescu, and S. Mihalache, "Particulate matter air pollutants forecasting using inductive learning approach," *Rev Chim-Bucharest*, vol. 67, no. 10, pp. 2075-2081, 2016.

[21] J. Dean, and S. Ghemawat, (2008). "MapReduce: simplified data processing on large clusters," *CACM*, vol. 51, no. 1, pp. 107-113. [Online]. Available: https://dl.acm.org/doi/10.1145/1327452.1327492

[22] I. Hashem, N. Anuar, A. Gani, I. Yaqoob, F. Xia, and S. Khan, "MapReduce: review and open challenges,". *Scientometrics*, vol. 109, no. 1, pp. 389-422, 2016. [Online]. Available: https://doi.org/10.1007/s11192-016-1945-y, https://link.springer.com/article/10.1007/s11192-016-1945-y?shared-article-renderer

[23] S. Shirwadkar, "An evaluation of key-value stores in scientific applications". MSc thesis, University of Houston, Texas, USA, 2017. [Online]. Available: https://uh-ir.tdl.org/bitstream/handle/10657/1864/SHIRWADKAR-THESIS-2017.pdf?sequence=1

[24] O. Siriporn, and S. Benjawan, "Anomaly detection and characterization to classify traffic anomalies case study: TOT public company limited network,". *WASET*, vol. 3, no. 1, pp.15-23, 2009. [Online]. Available: https://publications.waset.org/11934/anomaly-detection-and-characterization-to-classify-traffic-anomalies-case-study-tot-public-company-limited-network

[25] SAS, Predictive Analytics What it is and Why it Matters, 2017. [Online]. Available: https://www.sas.com/en_id/insights/analytics/predictive-analytics.html

[26] P. Chen, and C. Zhang, (2014). "Data intensive applications, challenges, techniques, and technologies: a survey on big data," *Inf. Sci.*, vol. 275, no. 1, pp. 314-347. [Online]. Available: https://scinapse.io/papers/2109574129

[27] P. Cota, D. Rodríguez, R. González-Castro, and M. Gonçalves, "Massive data visualization analysis of current visualization techniques and main challenges for the future. Proceedings of the Information Systems and Technologies," *IEEE 12th Iberian Conference* on June 21-24, 2017, pp. 190-195, Lisbon, Portugal. [Online]. Available: https://ieeexplore.ieee.org/document/7975704

[28] H. Teh, A. Kempa-Liehr, K. and Wang, "Sensor data quality: a systematic review," *J Big Data*, vol. 7, no. 11, pp. 1-49, 2020. [Online]. Available: https://doi.org/10.1186/s40537-020-0285-1, https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-0285-1

[29] Y. Alsultanny, "Data mining and visualization: meteorological parameters and gas concentration use case," Proceedings of the XIX International Conference on Data Analytics and Management in Data Intensive Domains. *DAMDID/RCDL'2017*, October 10-13, 2017, Moscow, Russia.m [Online]. Available: http://ceur-ws.org/Vol-2022/paper53.pdf

[30] X, Chen and X. Chen, "Data visualization in smart grid and low-carbon energy systems: A review," *International Transactions on Electrical Energy Systems*, vol. 31, no. 7, pp. 1-12, 2021. [Online]. Available: https://doi.org/10.1002/2050-7038.12889

[31] I. Maletic and A. Marcus, "Data cleansing: beyond integrity analysis," Proceedings of the *International Conference on Information Quality*, October 20-22, 2000, pp. 200-209, Massachusetts Institute of Technology, USA. [Online]. Available: http://mitiq.mit.edu/ICIQ/Documents/IQ%20Conference%202000/Papers/DataCleansingBeyondIntegrityAnalysis.pdf

[32] J. Vaske, J. Beaman, and C. Sponarski, "Rethinking internal consistency in Cronbach's Alpha," *Leis. Sci*. vol., 39 no. 2, pp. 163-173, 2017. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01490400.2015.1127189