Some Future Directions and Environmental Impacts of Air Pollution

Naser Ali Abdullah[1], Nishtiman Y Mosa[2*], Khonav IK Mamil[3], Dilan Jassim Khalil[4]

[1]Duhok polytechnic university; Amedi Technical Institute; Tourism management Department
[2*]University of Duhok, College of Health Science; Anesthesia Department.
[3]Ministry of Education ; Duhok directorate of Education
[4]College of Science, Sicentific Research Center, University Of Duhok

***Corresponding Author:-** Nishtiman Y Mosa
[*]University of Duhok, College of Health Science; Anesthesia Department. e-mail: Nishtiman.rashe@gmail.com

Abstract
Air quality physics are less relevant to data - driven air quality models than system identification theory . They are used specifically to define a broad range of mathematically calibrated causes-and-effect relationships between tuples of input and output data. Potential causes include the quantity of emissions of a particular group of pollutant precursors both inside and outside the research domain, local climatic information, and concentrations of a particular group of pollutants in earlier temporal steps. The expected concentration of a pollutant (or several pollutants) is one of the consequences that is often taken into account . These models' key benefits are their low computational resource requirements and ease of implementation . The amount of air pollution is increasing daily and has an impact on both human health and the ecosystem .
As a result , it's crucial to control it by maintaining a continual air check at its level. We create a model of multi..sensor data fusion with the ability to identify and predict the worst gas in order to lower the level of pollution . In this study , we offer an effective method for clustering the data from several sensors, which was formerly used to divide and categories the data .
In order to forecast air quality, this research suggests a random to assessed using actual data from several cities .

Keywords:- environment, air pollution , forecasting, nitrogen oxide emission, implementation system.

1 – Introduction

Contaminants in the air we breathe might be either organic or man – made . which is referred to as air pollution, internal and exterior air pollution are the two main categories that it is commonly categorized under . Exposures to un - built ecosystems result from external air pollution . For instance , the air is contaminated by pollution from factories, power plants and other sources, such as automobiles . Unlike indoor air pollution , which is brought on by a variety of things such home items , chemicals, building supplies and gases **(Yunliang *et al.*, 2016)**.

It has been possible to monitor the level of air pollution thanks to the air pollution control system . Numerous health impacts have been documented during the past 30 years in studies, and many of them are thought to be caused by exposure to air pollution . Some air contaminants are dangerous .

The likelihood of developing health issues can increase if you breathe in these contaminants **(Relvas *et al.*, 2017)** .
Children , teenagers in their late teens, additionally, people who already have heart or lung issues are more susceptible to the negative effects of air pollution. It does not follow that pollution only exists outside ; toxic air can also exist inside of buildings and be harmful to your health . A release of contaminants into the atmosphere is what produces air pollution , which is the word used to describe the presence of compounds in the air known as pollutants **(Kanjo *et al.*, 2018 ; Turrini *et al.*, 2019)** .

The makeup of the air is altered by these contaminants , making it dangerous . The combustion of fossil fuels is the main contributor to air pollution because when they are burned , a variety of chemical substances are released into the atmosphere . The effects of air pollution on people and the environment are significant **(WenxiuDing *et al.*, 2019)** .
Human respiratory and cardiovascular issues result from this . Humans' organs are impacted when they breathe in contaminated air. Other impacts include eutrophication, ozone depletion , industrial pollution , global warming and others . Utilizing solar and wind energy, energy-efficient appliances, public transportation and other methods can all help to reduce air pollution . Pollutants are released into the atmosphere less when people use public transportation as opposed to private transportation . Utilizing energy - efficient machinery and renewable energy sources , such as solar and geothermal energy , can reduce the quantity of power generated by burning fossil fuels (**Turrini *et al.*, 2019** ; **WenxiuDing *et al.*, 2019**) .

Data-fusion is the process that is for combining the usingdata from different sources to provide more reliable results . When several sensors' data are combined to generate descriptions of an environment , this process is known as sensor fusion . Information fusion is a subset of sensor fusion, which is referred to as "multi - sensor fusion" . There are two subcategories of multi - sensor data fusion: Homogeneous-data fusion and Heterogeneous-data fusion **(Carnevale *et al*., 2015; Jia *et al*., 2020)** . When data from one type of sensor are combined, it is called homogeneous data fusion ; when data from other types of sensors are combined , it is called heterogeneous data fusion .
Signal level fusion, object level fusion, function level fusion, and decision level fusion are a few of the layers of data fusion processing.

2..Materials and Methods:--
2..1. Data Collection
Over the past several years interest in Data Fusion (DF) technologies has grown because systems now produce a vast amount of data that needs to be combined. Significant applications , innovations, and demographics have different interpretations of the data consolidation . In a more general sense , In order to determine the characteristics of an ecosystem or other entity, data from a variety of sensors must be gathered and analyzed it . Comparing data from multiple sources with DF reveals significant advantages . Besides the statistical advantage of merging data from the same source .

The accuracy with which a volume can be measured and classified will be improved by making use of the data from these numerous sensors . This study uses data on air pollution that is collected in real-time from various air quality monitoring sites . Gathered the meteorological information , such as the pressure, theRelative humidity, air temperature, wind speed, and wind direction all have an impact on the concentrations of air pollutants.
There is a function for sensor fusion in this . As an example , a particle sensor can be used to identify physical contaminants while other sensors are utilized to capture data during the data collection process **(Constantinescu *et al*., 2007; Skamarock *et al*., 2008)** .

Thermal sensors are used to gauge temperature and humidity. Utilizing gas sensors , gases such as O_2 , NO_2 , SO_2 , CO_2 and CO are observed .
Heterogeneous data fusion is used to combine various sorts of data from various types of sensors after the data have been collected , combine the three different types of location information into one . The original data , however , could occasionally have duplicate records or incomplete values that are missing . The collected time series needed to be revised and updated as a result of the numerous fluctuations and scenarios that were present and caused noise in the timelines **(Hamed *et al*., 2019)**.
Preprocessed data must therefore be clustered. or categorized in accordance with the fusion principles before being filtered .
For clustering , it was proposed an LDA using a random forest tree following classification (A method of dimensionality reduction is LDA . It is employed to decrease the dimensions .
It is used to identify group differences and to divide at least two classes. LDA One of the clustering techniques is used to collect and separate data;and when compared to other algorithms , it delivers the best accuracy . The same dataset must be subjected to LDA twice , with each application utilizing a different job . LDA will serve as a classifier in the primary methodology and in the secondary methodology , it will lower the dataset's dimensionality and The grouping task will be handled by a neural system. The results of the two methods will then be compared. **(Punyasha, 2018)** .
In cases where the univariate (single knowledge variable) only has two groupings) . By dividing the total number of characteristics by their aggregate total , the usual route can be used to obtain the Estimates for each class (ki) and the mean (m) estimate of each piece of information (xi). m = 1/nki*sum(xi)
About the class K;the total number of occurrences of class k, or nki , is given where m is an estimated average value for x.. The adjustment can be determined by subtracting each reward's value from the average across all groups as determined by conventional squaring.
$sig^2 = 1/ (n – Ki)* sum((xi – m) 2)$
Where m is the average of the inputs (xi), n is the number of instances, Ki is the number of groups, and sig2 is the total variance.
By calculating the probability that each input set will belong to a particular group , LDA performs the prediction process .
The output class is determined by the group with the highest probability and a prediction is then given for that group.
To calculate probabilities , this model applies the Bayes Theorem . Using the probability of each class It is used to calculate the probabilities of the given output class (ki), the input class (xi), and the output class, as well as the likelihood of each class's data. Given an output class (ki) and an input (xi), calculate P(Y=x|iX=xi) as follows: (Pok * fki(xi))/ sum (PIl * fki(xi))
Where Pok denotes the fundamental likelihood of each class (ki) discovered after training . This is referred to as the prior probability in the Bayes' theorem .
Pok = nki/n
The possibility of xi falling into a specific class is represented by the following formula, f(xi). A Gaussian distribution FUNCTION is utilized for that function f (xi) , that obtain by resolving the previous equation ,

D (xi) = xi * (m/sig$_2$) – (m^2/(2*sig^2) + ln (Pok)
D (xi) The discriminating function for class ki with input xi is where they all, sig2 and Pok, come from.
Better results can only be obtained using LDA as a classifier . Extending from the previous point , Although there Due to non-linear data features, class duplication may still occur in some cases., the LDA can make every effort to reorganize those data set in a new gape to obtain a greatest feasible linear separability if the dataset is no longer linearly separable . Two examples of classification techniques you may use in this case to deal with nonlinear data are a neural network model with hidden layers and a neural network with accessible radial basis functions .
(Dixian *et al.*, 2018; Punyasha, 2018) .

2 – 2 – Random Forest
This computation is one of those for managed arrangements.
The large dataset is organized using this system . A choice calculation is the strategy the organization employs to deliver a precise projection . In order to "realize" how to organize unlabeled data , a controlled learning model that makes use of labeled data is the irregular timberland calculation . The K-implies Cluster computation is at odds with this. It organizes the vast dataset with enormous size . The method the organization uses to provide an accurate forecast is a choice calculation. Unreliable calculation for timberland is a controlled learning model that makes use of tagged data to "realize" how to organize unlabeled data . The K - implies Cluster computation is at odds with this **(Punyasha, 2018 ; Hamed *et al.*, 2019**) .

Using this method , it can figure out which branches are more likely to appear on each node for each branch by using the class and probability . In this dataset , pi represents the class frequency, and c indicates how many classes there are overall. If implemented properly , several different types of data sets would benefit from the random forest method , including those used for classification or regression analysis . It is easy to use , quick to train and accurately depicts the decision trees it employs .

2 – 3 – Data Analysis
Prior to dividing the remaining documents into time segments , the major goal of the data pre - processing is to remove duplicate documents from the data . We will add phrases to the alias list and the alias will take effect when terms with the same meaning are found in the clustering network . Additionally , Microsoft Excel 2019 allows you to export the relevant data for de –duplication .
The components of the adopted technique and their relationships were depicted in Figure 1; the system calculates forecasting of P M 10 conc. at time t $PM10R(t+1,\ldots,t+n)$ at-time ($t+1,\ldots,t+n$) using which: the results of the re-analysis ..phase which integrates the output of the Comprehensive *Air Quality model* with extensions (CAMx) model ($PM10G(t)$) and the measurements .

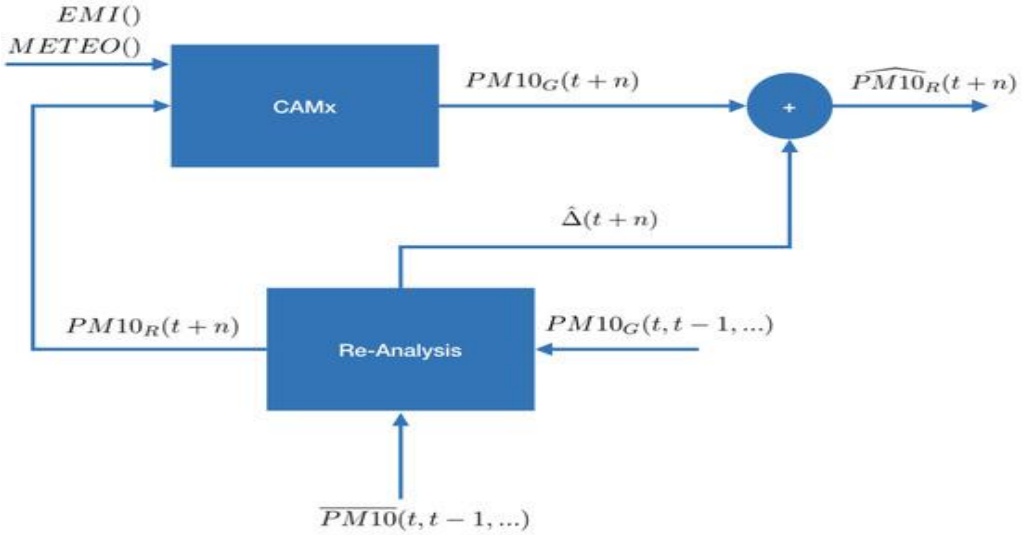


Figure 1 . The system's implementation in a diagram

3 – Results & Discussion
Regional and municipal authorities rely heavily on deterministic forecasting models for air quality since they are essential instruments for assuring the timely delivery of information on current or potential pollutant value excesses . Most of the time, these models overestimate some important pollutants, such P M 10 , specifically in areas of the high conc. , which is one of their fundamental flaws **(Jeon *et al.*, 2017)** . This makes it difficult to foresee critical occurrences , which are those that surpass the daily PM10 concentration limit of 50 g/m3 .
Rapid computational strategies are described , put into practice and assessed in this study in order to solve this issue .

The techniques are based on off-line correction of the prediction window output from the chemical transport model. , approximated by the data that measured up until the prediction window's start **(Environ. CAMx, 2020)** .
In particular , The approaches are based on a correction estimate that is displayed as a linear collection of corrections that were calculated for the days where measurements are available .
To obtain predictions with a high degree of accuracy, real-time data from many locations and times is gathered . Locations are identified as hazardous based on the meteorological conditions and conditions that are the most severe after data from the datasets has been analyzed **(Neal *et al.*, 2014)** .
The suggested approach can forecast gas levels by combining sensor data from various cities, it is able to predict pollution. levels of various Gases by utilizing LDA and a random forest tree . The complete annual emission (t / yr) over the CAMx domain for the calendar year is shown in Table 1 lists each pollutant and the CORINAIR macro sectors for air emissions.Table 1 . over the CAMx domain, the total yearly emission (t/y r) of a given year.

(CORINAIR)Macrosector	(VOC)	(NH3)	No x	PM 10	PM 2.5	S O2
The burning of fuels in the energy and transformation industries	2017	54	31,573	840	717	8642
combustion facilities not used by industry	71,187	1081	50,167	42,544	51,248	701
Burning in the manufacturing sector	7203	864	70,022	4715	3025	26431
distribution of geothermal and fossil fuel energy	27,126	1	457	90	68	1362
Other portable equipment and sources	16,308	21	67,185	7246	6002	4710
Treatment and disposal of waste	3281	3710	5938	2035	1148	1409

Comparisons between the system's results and the daily PM10 readings taken by the monitoring stations have been made to assess the performances . There have been two phases to the validation process . The normalized mean absolute error (NMAE) and the correlation coefficient , two separate statistical indices , have been used to evaluate the forecasts in the first case. The second phase looked at the model's ability to accurately replicate the significant occurrences , they were characterised by values greater than the 50 g/m3 cutoff .
In Figures 2 , 3 and 4 , it can see a boxplot of the correlation coefficient for all the tests that were performed .
It is clear that the integration strategies offer superior outcomes of each the one-step-ahead (forecast) compared to a "standard" scenario (CAMx) .
In general , the LS test has higher coefficient values , however using a memory with a memory size of three did not improve performance.

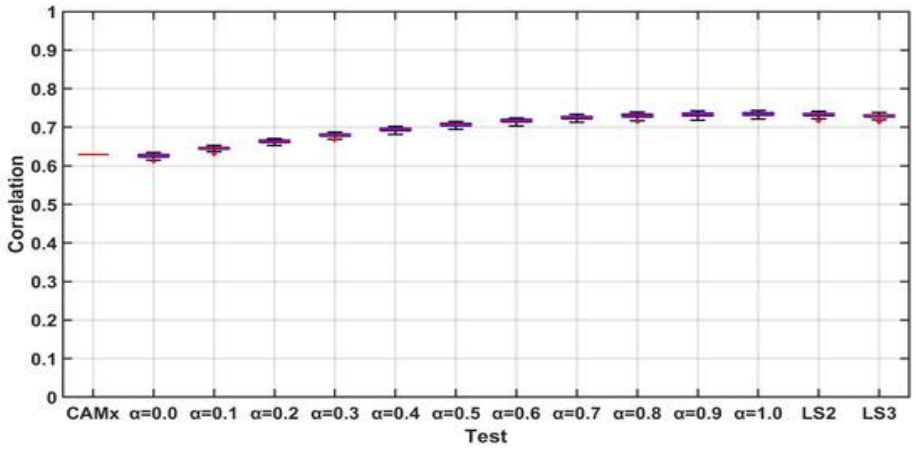
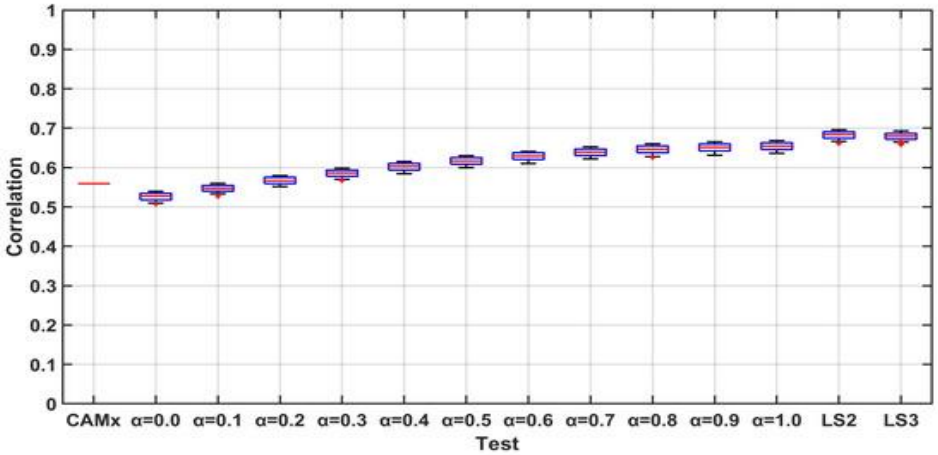


Figure 2 . For the single-step forecast, there is a forecast correlation.



(Figure3). Forecasting in two steps with correlation.

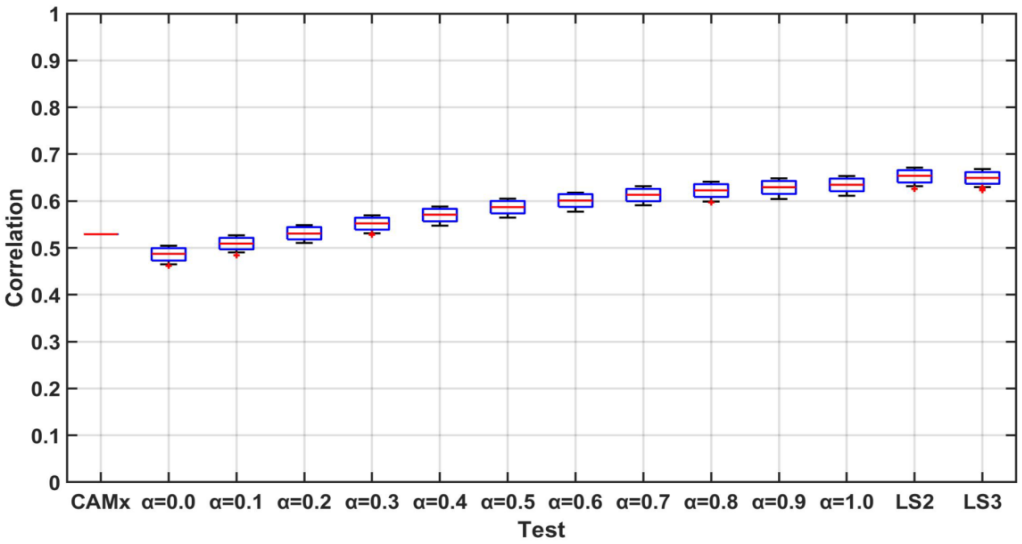


Figure 4 . Forecasting in three steps with correlation

4 – Conclusion

In this study , two methods for integrating measurement data and model results for forecasting applications are presented .The techniques rely on the results of the best interpolation between measurements and models up to the start of the prediction to determine the adjustment that needs to be added to the CTM results in the forecasting horizon.. There have been tests of two various approaches :

(i) Using the most recent two calculations of corrections ,

a simple weighted mean technique is used .

(ii) A more advanced least-square error method is used to connect the most recent rectification computations by estimating the linear regression coefficient .

The computation is quick in both of these examples and the CTM model code does not need to be altered . To determine the concentration of PM10 , the methodology has been integrated on a CAMx - based system . For the computation of PM10 concentration , the methodology has been integrated on a CAMx -based system **(Environ. CAMx Version 6.50)** .

The performance of the integrated system decreases a little bit more quickly than the chemical transport model , according to a sensitivity analysis of the prediction skill with regard to the forecasting horizon , most likely because meteorological data forecasting was not taken into account . In terms of the choice of the potential places for measurement station placements , the assessment demonstrates the methodology's reliability .

These factors together with the results of the integration methodology seem to produce a system that the Local Authorities can use to foresee significant occurrences involving PM10 in a comprehensive manner . Although there are some unique characteristics of the examined pollutants that must be taken into account , most notably the influence of very localized nitrogen oxide emissions , in spite of this , at other dimensions and in different domains , Other primary and secondary pollutants, like ozone and nitrogen oxides, can also be analyzed using this method.

References

1. Carnevale, C.; Finzi, G.; Pederzoli, A.; Pisoni, E.; Thunis, P.; Turrini, E.; Volta, M. A methodology for the evaluation of re-analyzed PM 10 concentration fields: A case study over the PO Valley. *Air Qual. Atmos. Health* 2015, *8*, 533–544
2. Constantinescu, E.M.; Sandu, A.; Chai, T.; Carmichael, G.R. Assessment of ensemble-based chemical data assimilation in an idealized setting. *Atmos. Environ.* 2007, *41*, 18–36
3. Environ. CAMx User's Guide Version 6.50. Available online: http://www.camx.com/files/camxusersguide_v6-50.pdf (accessed on 8 January 2020).
4. - Dixian Zhu, Changjie Cai, Tianbao Yang and Xun Zhou, A Machine Learning Approach for Air Quality P rediction: Model Regularization and Optimization , Big Data Cogn. Comput. 2018, 2 , 5; doi: 10. 3390/ bdcc2010005
5. Hamed Karimian, Qi Li, Chunlin Wu, Yanlin Qi, Yuqin Mo, Gong Chen, Xianfeng Zhan g, Sonali Sachdeva, Evaluation of Different Machine Learning Approaches to Forecasting PM 2.5 Mass Concentrations, Aerosol and Air Quality Research, 19: 1400–1410, 2019
6. Jeon g-Joo Kim, Su-il Choi, " Implement at ion of Pothole Detection System Using 2D LiDAR", International Conference on Electronics, In format ion , and Communication, 2017 , 607 -610.
7. Jia Liu, Tianrui Li, Peng Xie, Shengdong Du, Fei Teng , Xin Yang, Urban big data fusion based on deep learning:
8. An overview, Information Fu sio n 53 (2020) 1 23 –133

9. Neal, L.; Agnew, P.; Moseley, S.; Ordonez, C.; Savage, N.; Tilbee, M. Application of a statistical post-processing technique to a gridded, operational, air quality forecast. *Atmos. Environ.* 2014, *98*, 385–393
10. Punyasha Chatterjee, Workshops ICDCN '18: Air Pollution Detection Using Multi sensor Data Fusion Proceedings of the Workshop Program of the 1 9t h International Conference on Distributed Computing and Networking January 2018 Article No.24 Pages 1–2
11. - Kanjo a, Eman M.G. Younis b, Nasser Sherkat a, Towards unraveling the relationship between on -body, environmental and emotion data using sensor information fusion approach, Information Fusion 40(2018) pp 18 - 31
12. Relvas, H.; Miranda, A.; Carnevale, C.; Maffeis, G.; Turrini, E.; Volta, M. Optimal air quality policies and health: A multi-objective nonlinear approach. Environ. Sci. Pollut. Res. 2017, 24, 13687–13699.
13. Skamarock, W.; Klemp, J.; Dudhia, J.; Gill, D.; Barker, D.; Duda, M.; Huang, X.; Wang, W.; Powers, J. *A Description of the Advanced Research WRF Version 3: NCAR/TN-475*; National Center for Atmospheric Research: Boulder, CO, USA, 2008; Volume 88.
14. Turrini, E.; Vlachokostas, C.; Volta, M. Combining a Multi-Objective Approach and Multi-Criteria Decision Analysis to Include the Socio-Economic Dimension in an Air Quality Management Problem. Atmosphere 2019, 10, 381.
15. Wang, Jianhua & Ogawa, Susumu. (2 015). Effects of
16. Meteorological Conditions on PM2.5 Concentration s in Nagasaki, Japan. International journal of environmental research and public health. 12. 9089-101. 10.3390/ijerph120809089.
17. WenxiuDing a , Xuyang Jing a , Zheng Yan , Laurence T. Yang c , A survey on data fusion in internet of things: Towards secure and privacy-preserving fusion, Information Fusion 5 1 (2019) 129 –144
18. - Yunliang Chen, Lizhe Wang , Fangyuan Li , Bo Du , Kim-Kwang Raymond Choo , Houcine Hassan, Wenjian Qin, Air quality data clustering using EPLS meth od R, In format ion Fusion(2016) .