International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Using Generalized Additive Models (GAMs) for Large Datasets to Determine the Effect of Air Pollution on Human Health

A. O. Ochugboju[1], A.Yawe[2], H. A. Odiniya[3] and A. A. Musa[4]

[1]Department of Mathematics, Federal University Lafia, P.M.B. 146, Lafia, Nasarawa State Nigeria

[2]Department of Mathematics and Statistics, Federal University Wukari, P.M.B. 2010, Wukari, Taraba State Nigeria

[3]Information and Communication Technology Department Federal University Wukari, P.M.B. 2010, Wukari, Taraba State Nigeria

[4]Department of Economics, School of Art and Social Science, Umar Suleiman College of Education, P.M.B. 02, Gashua, Yobe State Nigeria

Abstract: ***The impact of air pollution and human health has been the subject of a colossal measure of epidemiological activity. This research considers an application in the effects of air pollution and human health where generalized additive models (GAMs) are proper. However, due to the large size of data, the use of GAMs is practically inflexible with existing methods. In this manner, Wood et al. (2014) developed generalized additive model fitting methods for substantial data sets for the situation in which the smooth terms are replaced by using penalized cubic regression splines. We extended our analysis to ten (10) cities simultaneously using generalized additive model. Through the utilization of the environmental package from the National Morbidity, Mortality, Air Pollution study, GAMs ends up being adaptable with large data set.***

Keywords: Air Pollution, Generalized Additive Models, epidemiological activity, cubic Regression Splines

1. Introduction

Regression model containing several observational variables has become cumbersome for existing statistical tool. The reason is due to large data size that have made their use practically inflexible. This research considered a current model known as Generalized Additive Model (GAMs) to be practical in taking care of this huge data issue through the utilization of the recent statistical software. One example of these massive data is the National Morbidity, Mortality and Air Pollution Study data (NMMAPSdata). The effect of air pollution on human health will be our focus of study. The package that is compiled by Peng et al. (2004) contains a daily mortality rate of 108 cities in the United States. According to Peng et al. (2004), the study has added consistent evidence of the severe health effect of particulate matter.More so, the Environmental Protection Agency (1996) added that the fundamental part of the NMMAPS in the improvement of the air quality measures pulled in substantial analysis from mainstream researchers and industry gatherings. The analysis is with respect to the statistical models that are applied and the techniques used for adjusting for potential confounding Peng et al. (2004).

The study made use of the National Morbidity Mortality Air Pollution data package. It is data compiled by Peng et al. (2004) for 108 cities in the United States of America over a period of 10 years. However, our interest is in ten cities: Chicago, Akron, Atlanta, Austin, Bakersfield, Baltimore, Baton Rouge, Cincinnati, Clevland, and Corpus Christi. This research considered these cities to see if these cities could be modeled simultaneously using GAM.

2. Literature Review

Several findings have shown that particles matter (PM) with an aerodynamic diameter less than 2.5m have the strong relationship with respiratory-related death and disease than coarse particles (Wichmann 2007). However, the levels of evaluations on adverse health impacts were conflicting crosswise over many countries and areas. As a result, there is a little confirmation to propose a limit below which no adverse health effects would foresee (Peng et al., 2004). Furthermore, Peng et al. (2004) stated that the numerous studies have played a significant part in setting the right standard for the ambient particulate matter. Examples of these multicity studies are the National Morbidity, Mortality and Air Pollution Study (NMMAPS), the Air pollution and health: a European approach study and analyses of Canadian cities. Our interest in this researchwould be on the NMMAPS data.

According to Katsouyanni and Samet (2009), the aim of these larger analysis has been to create more dependable estimates of the potential severe effects of air pollution on human health. Another aim is to provide a common ground for a contrast of risks across geographic areas. Thirdly to increase the ability to differentiate between health effects associated with air pollution and those linked to other factors. Ultimately, the goal is to improve the scientific basis for decisions about whether and how to regulate air pollution (Katsouyanni and Samet, 2009). However, multicity time-series studies of particulate matter and mortality and morbidity have shown strong evidence of a high relationship between daily environmental air pollution and daily mortality(Peng et al., 2004). This evidence has formed an enormous epidemiological review of the US national

ambient air quality standard for particulate matter. Multicity time series studies use data on a day by day concentration of air pollutants as well as daily measures of health impact primarily at the level of a single city (see examples Pope et al. (1995), Bell and Ebisu (2008), Pereira et al. (2014)). Taking into consideration death count and number of admissions to hospitals. However, the wide range of methods used to assemble and scrutinize data from individual cities has made their findings difficult to interpret(Samoli et al., 2008). Hence, has led to advanced efforts to combine information across multiple cities and ultimately, across geographic regions (Katsouyanni and Samet, 2009). According to Daniels et al. (2000), time series studies on pollution and morbidity are analysed by using log-linear distribution, Poisson regression models for overdispersed counts with the frequent count of deaths recorded. From the World Health Organisation (WHO 2006), the global burden of disease study estimates that urban particulate matter exposure was the cause of nearly 62,000 lung cancer death in the year 2000. There are an expansive number of time series studies that evaluate the fleeting health impacts of fine particle pollution. These studies are of particular importance in mortality and morbidity, by a method of fitting Poissonregression models at a group level (Katsouyanni and Samet, 2009). In a survey carried out by (Wichmann et al., 2000) in Erfurt, Germany, demonstrated to some degree sturdier relationship with respiratory diseases as regards to cardiovascular disease and other effects. Moreso, the levels of the assessments of unfavorable health impacts were conflicting crosswise over numerous nations and areas, and there was little confirmation to recommend a limit underneath that no antagonistic health impacts would be expected. Reports in various fields have demonstrated that the attributes of the particle mixture change during the time. And also the relative and outright commitments of specific components to particle mass may be diverse amid distinctive periods of the year (Peng et al. (2004), Bell and Ebisu (2008)). According to Dominic et al. (2006), other potential time-changing confounding and altering components, for example, temperature and humidity can likewise influence evaluations of transient effects of particles on respiratory mortality and morbidity contrastingly in diverse seasons. Time series investigations of the impact of both particles and weather conditions on mortality have distinguished the significance of satisfactory control for temperature and humidity when evaluating air pollution effects (Samet et al., 2000).

3. Data

3.1 Data Description

The NMMAPS database is package compiled by Peng et al. (2004). It is made up of daily mortality/morbidity counts, air pollution levels, and weather variables. According to Peng et al. (2004), the NMMAPSdata package was first collected for the National Morbidity, Mortality, and Air Pollution Study (NMMAPS). A day to day mortality records was gotten from the National Center for Health Statistics and classified into three age categories (Samet et al., 2000). The age groups are group A, under 65; group B, 65-74; group C from 75 above. According to Samet et al. (2000), accidental death counts were omitted. While the weather data were received from the National Climate Data Center EarthInfo CD-ROM and the air pollution levels were obtained from the Environmental Protection Agencies Aerometric InformationRetrieval System (AIRS) and AirData System. It is important to note that hospital admissions were not included in the package. The data is used as a tool for epidemiological analysis, investigating the impact of air pollution/mortality of 108 U.S citiesover a period of fourteen years (1987-2000) (Peng et al., 2004). Thus gives a totalof 5,114 days of observation for each of the age groups. For instance, Chicago cityhas a total of 5,114 days of observations. Since the age category are three (3), we, therefore, have a total of 15,342 observations. A similar calculation follows for the rest of the cities. The data are distributed into 108 distinct data frames giving a total of 15,342 rows and 291 columns. In expansion to giving the NMMAPS database as a single element, it incorporates capacities for building ”versions” of the database that may be more suitable for various types of analyses. The package is a structure for running well-ordered, efficient, and reproducible analyses of time series data on air pollution and mortality. With the requirement for reproducible exploration in epidemiologic studies just expanding, the package has been intended to encourage and empower such reproducible analyses. Moreover, the package streamlines dispersion of the information and makes a typical stage for scattering results and procedure (Peng et al., 2004). It is important that while NMMAPS is at present the biggest database connecting day by day mortality with air pollution exposures, it is little contrasted with datasets regular to fields, for example, genomics or remote sensing.

4. Methodology

4.1 Models for Large Datasets

Wood el al. (2014) explains another example how Generalized Additive Model (GAM) can be made feasible for large data set, taking the relationship between air pollution and human health. This new type of analysis is now being used for large air pollution epidemiology studies. It makes use of the NMMAPSdata assembled by Peng et al. (2004) containing 108 cities over a period of years. Each city has a mortality rate due to respiratory issue according to their age category. One of the ways used by Wood (2006) was to try to model the mortality rate related to respiratory issues, at the level of a single city. This approach is regarding pollution and climate variables as well asthe length of time in contextual death term. The data which are enclosed in a dataframe depending on a particular city concentrated on the day to day mortality rateover a period of years. According to Wood et al. (2014), the model for Chicago propose an extremely high relationship with ozone and temperature when the variablesare accumulated over the four days up to and including the day of death. The causeof the death rate was due to levels of ozone, levels of sulfur dioxide, mean daily temperature, and levels of particulate matter. Moreso, Wood et al. (2014) considers themodel to be intriguing as it fits various days of very strong death rate that generallywould be huge anomalies. However, Wood et al. (2014) is interested in finding out ifsuch high interaction is constant with the rest of the 108 cities. Also to find out if thewhole dataset assembled by Peng et al. (2004) could be model simultaneously as inthe case of the Electricity grid

load predictionWood et al. (2014) using the model below:

$$\log(E[Y_i]) = \beta_{j(i)} + \alpha_{k(i)} + f_{k(i)}(x, z) + f_4 \quad (1)$$

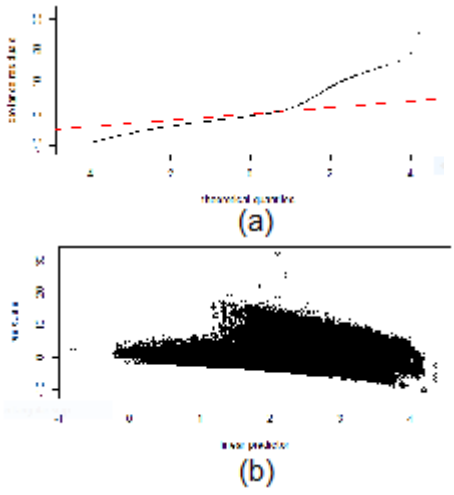
where, is the observation, that is mortality rate of a particular city say and represents the age group running from group 1 - 3. represents the daily time, represents, ozone having zero mean and z represents the temperature (F). Thetemperature and ozone are both added over the four days with the day of deathinclusive. The mortality rate is assumed to follow a Poisson distribution. is a smooth function, as such to can be modeled by tensor products of cubic regressionalsplines as there are 5114 observations(Wood et al., 2014). Each of the tensor productshas a basis dimension of 100 each and, of course; all smooth were penalized for eachdimension. In addition to the air quality variables, the basic mortality rate seems tovary with time. Wood et al. (2014) modeled f4 using a cubic regression spline basis ofdimension 400. In spite of the fact that the rearrangements of a single smooth of timefor all the 108 cities is fairly crude, the deviance based estimates of the scale parameteris just 1.03, proposing almost no overdispersion here. From his findings, it shows upto a great extent sensible, with the exception of the huge outliers comparing to theChicago outliers. In the case of investigating Chicago alone, these outliers vanishas a consequence of the ozone-temperature interaction. In any case, as indicatedby Wood et al. (2014), the full model estimates of these impacts recommends thatthis interaction alone is not adequate to clarify the Chicago outliers. Accordingly,the impacts are much weaker at high ozone-high temperature when contrasted withthat found in Wood (2005, section 5.3), thus recommending overfit to Chicago alone.With identifiability constraints, the model had 802 coefficients and was evaluatedutilizing 1,210,113 samples. The model matrix alone would require in abundance of7Gb of capacity in the event that it was generated full. From the Choleski type ofthe technique, a 5% testing way to get initial values, model setup and estimationtook 12.5 minutes on a machine with a 3.1Ghz Intel i3 540 processor and 3.7 GbRAM, running Linux (i.e. a PC retailing at not exactly 600 USD) (Wood et al., 2014).

Likewise, Wood utilized a 4 fold parallelization strategy, despite the fact that since 2of the processors 4 centers are virtual the technique with 2 fold parallelization takesjust 2 minutes longer Utilizing the more steady QR approach more or less multipliedfitting time. A single threaded ATLAS BLAS Whaley et al. (2001) was utilized (thenon-parallel adaptation of the technique, utilizing a multi-threaded ATLAS BLAS,about doubled the calculation time). The model grid estimation plainly makes thismodel infeasibly broad for the Wood (2011) procedure executed in routine gam ofpackage mgcv. Regardless it should be seen that the Wood (2011) frameworks takes11.5 minutes to assess the same model to just the 0.9% of the information that startfrom Chicago, with a memory footmark of around 1Gb. Given that the Wood (2011)computational cost is linear in issue size, the new framework in this way offers a 100 foldrate for this issue. The backfitting approach to managing GAM estimation (Hastiesand Tibshirani,1986,1990), as realized in R package gam has to some degree lowermemory necessities, however without smoothness estimation. Package gam cannotfit absolutely the desired smooth-age class interaction, yet in any case it exhaustedall structure memorywhen trying to fit a streamlined variation of the model with asingle ozone temperature association for each of the three age classification to the fulldataset. The same was correct for the Iterated Nested Laplace Approximation of Rueet al (2009).

5. Air Pollution in Ten (10) Cities

We would now take ten cities from the National Morbidity Mortality Air Pollution data and run gam. The cities are Chicago, Akron, Atlanta, Austin, Bakersfield, Baltimore, Baton Rouge, Cincinnati, Clevland and Corpus Christi. Since there 15,342 observations for a city, a total of 153,342 observations and 291 variables will now be gotten from the combination. Assuming the observed number of deaths in these cities have Poisson distribution with an underlying mean, that is the multiplication of a basic, time varying, death rate and the product of ozone and temperature as the dependent effect. Using model (2) for the total ten Cities data gives figure 1 below:

Figure 1 represent a model checking plot for the ten cities. Plot (a) represent Q-Q plot in model (2), we see the four outliers suggesting a very strong interaction of ozone and temperature over the 4 days up to the day of death. However, most of the plots lie outside the normal indicating that the plot is not normal. Plot (b) is the residual plot verses the linear predictor. In this case, the predictor variables are the ozone, particulate matter, sulphur dioxide and temperature. Since Generalized additive model is a penalized Generalized linear model where the smoothfunctionsare being replaced with some linear terms. It is important to check residual plots.Each point represents a day observation and giving a densely packed plot. That isthere are 5,114 days of observation for each city, thus there will be a total of 153,342days of observations. Furthermore, most of the residual points are positive showingthat the data points are above the regression line. One could also see the four outliersfound for Chicago city alone also appearing for the ten cities put together. In plot (c),we have the histogram of the residual plot. It shows the distribution of the residualsfor all observations. Finally plot (d) gives response vs. fittedvalues. It shows thatthe randomly distributed response have constant variance. There are also four outliersseparate from the other observations.



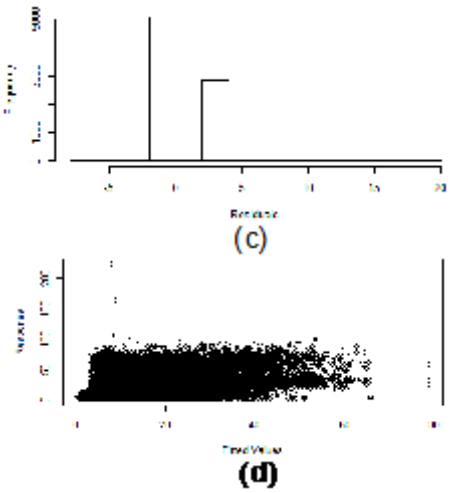


Figure 1: A model checking plot for model (2) for 10 cities. All the plots make clear that there a few outliers as seen in the previous checking plot. However from the Q-Q plot in (a), it shows that the plot is not normal as most of the points lie outside the regression line

Replacing the linear terms with smooth function using model (3) above gives a better fitting. Find a representation of the plots below

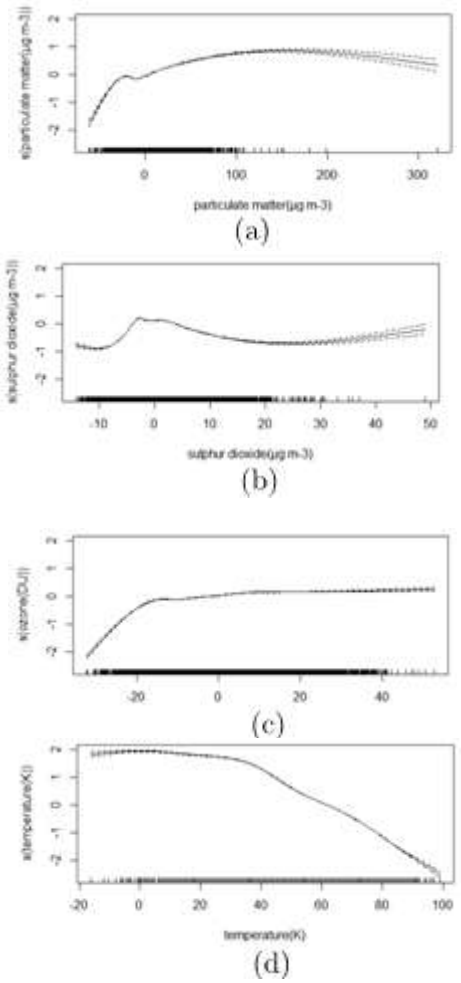


Figure 2: This is an assessment of the smooth from model (3). It is shown without partial residuals.

This method fits a model having several multiple predictors by constantly updating the fit for each predictor in turn, while keeping the other predictors fixed. However, the fitting for the ten cities is better using smooth function. It also shows that while keeping other pollutants constant temperature and ozone tend to have high interaction as seen in plot (c) and (d) respectively.

Now, if we consider checking a plot of date vs death of the entire ten (10) cities, the plot would be seen below:

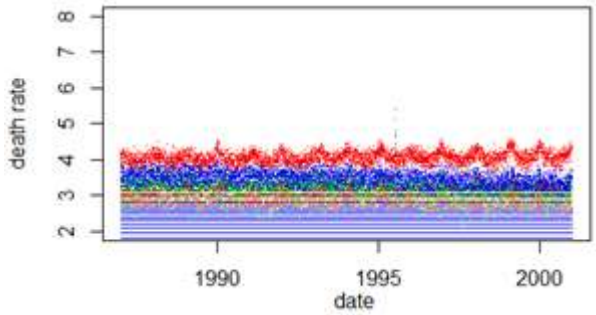


Figure 3: It shows the date rate vs death rate of the three age category. The red colour is the age category of 75 and over, the green colour is the age category of 65-74 and the third age category is those under 65 with colour blue.

Figure 3 above gives the plot of date vs. death rate of the ten cities with the way the different age group were affected. There is an indication that the higher age group that is, the group 75 and above have higher mortality rate as compared to the other 2 age groups. It is represented in red. While the next group represented in blue are within the age group under 65 and are the next group highly affected. Finally group represented green are less affected as compared to the other two and they are the group of 65-74.

We could also plot temperature vs. death rate to see the level of effect it has on the three-age category. Figure 10 shows this effect.

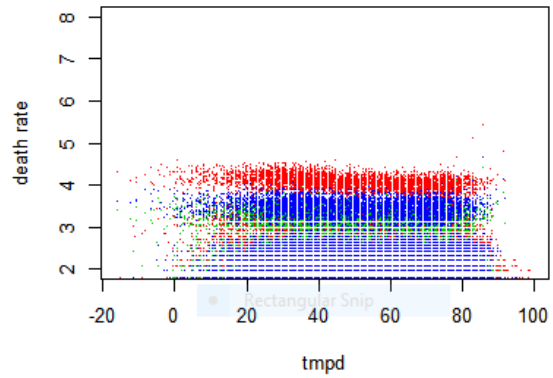


Figure 4: It shows the temperature rate vs. death rate of the three age category. The red colour is the age category of 75 and over, the green colour is the age category of 65-74 and the third age category is those under 65 with colour blue

Figure 10 above is the plot of temperature against death rate. From the plot, age category three tends to be highly affected. This category of people are those over the age of 75. It shows that as temperature increases, more of the people in this age group tends to be affected. It follows that high exposure to temperature as one gets older could lead to death. The next age group affected is the group under 65. This age group are the more active and stronger people who are mostly under the sun working. Also the children fall into

this category. Finally, the age group between 65-74 are the next.

6. Conclusion

Generalized additive models are practically direct tools that permit one to integrate nonlinear forms of predictor impacts into their linear models. Moreover, they permit one to stay inside of the already accustomed linear and generalized linear forms, while giving new ways of models investigation and potentially enhanced results. This process is very useful as it can be carried out simply by extending the existing methods and at the same time giving better accuracy and speed in the model fitting. The implication of this method is that we do not need to manually try out different computations on each variable individually (James et al., 2013). Generalized additive models prove to be more flexible when it comes to the higher order of polynomial regression models. One essence of GAM is that one can use these methods as building blocks for fitting an additive model. Our findings show that there is a strong relation between air pollution and human health. It is important to note that the observed death rate y follows a Poisson log distribution with mean zero which is related to the model predictors through a link function. The main idea is that the linear predictors now incorporate smooth functions f of at least some if not all the covariates. Obviously the idea that few days of high ozone and temperature can bring about increased mortality rates can clarify these data in the Ten Cities. To see the rate of mortality for each city, we would calculate the summary statistics for the ten (10) by finding their Standard Deviation, Mean, and Variance. The following table gives the result.

Table 1: Summary Statistics for ten (10) Cities

Number	City	Standard deviation	Mean	Variance
1	Chicago	15.8913	38.473	252.533
2	Akron	2.691	4.240	7.247
3	Atlanta	3.712	7.696	13.780
4	Austin	1.756	2.863	3.084
5	Bakersfield	2.078	3.367	4.318
6	Baltimore	3.088	6.709	9.536
7	Baton Rouge	1.534	2.517	2.370
8	Cincinnati	4.226	6.866	17.858
9	Clevland	6.692	12.472	44.789
10	Corpus Christi	1.322	1.966	1.747

From the table above, it is clear that Chicago have the highest number of mortality rate of 15.891 for the period under review. This is followed by Clevland with a standard deviation of 6.692. However the total ten Cities gave a standard deviation of 11.968, a mean of 8.716 and variance of 143.235.

References

[1] Bell, M. and Ebisu, K. (2008). Air pollution and birth weight. Environ Health Perspect,116 (3).
[2] Cao, J., Xu, Q., Chen, B., and Kan, H. (2012). Fine particulate matter constituents and cardiopulmonary mortality in a heavily polluted chinese city. Environ Health Perspect, 120.
[3] Daniels, M., Francesca, D., Samet, J., and Zeger, S. (2000). Estimating particulate matter-mortality dose-response curves and threshold levels: an analysis of daily time-series for 20 largest us cities. Am J Epidemiol, 152(5).
[4] Dominic, F., Peng, R., Bell, M., Pham, L., Dermott, M., Zeger, S., and Samet, J, M.(2006). Fine particulate air pollution and hospital admission for cardiovascularandrespiratory diseases. J Am Med Assoc, 295(10).
[5] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction toStatistical Learning. Springer, New York.
[6] Katsouyanni, K. and Samet, J. (2009). Air pollution and health: A european andnorth american approach (aphena). Health Effects Institute, 142.
[7] Peng, R., Welty, L., and McDermott, A. (2004). The national morbidity, mortality,andair pollution study database in r. John Hopkins University, Dept. of BiostatisticsWorking papers, 44.
[8] Pereira, G., Belanger, K., Ebisu, K., and Bell, M. (2014). Fine particulate matterand risk of preterm birth in connecticut in 2000-2006: a longitudinal study. Am JEpidemiol, 179(1).
[9] Pope, C., Thun, M., Namboodiri, M., Dockery, D., Evans, J., Speizer, F., and Heath,C. (1995). Particulate air pollution as a predictor of mortality in a prospective studyof u.s. adults. American Journal of Respiratory and Critical Care Medicine, 15.
[10] Samet, J., Francesca, D., Zeger, S., and Dockery, D. (2000). The national morbidity,mortality, and air pollution study,part i: Methods and methodologyic issues.
[11] Samoli, E., Peng, R., Ramsay, T., Pipikou, M., Touloumi, G., Dominici, F., Burnett, R., Cohen, A., Krewski, D., Samet, J., and Katsouyanni, K. (2008). Acute effects of ambient particulate matter on mortality in europe and north america: Results from the aphena study. Environ Health Perspect, 116(11).
[12] Whaley, R. C., Petitet, A., and Dongarra, J. (2001). Automated empirical optimizations of soft-ware and the atlas project. parallel computing. Elsevier, 27.
[13] Wichmann, H., Spix, C., Tuch, T., Wolke, G., Peters, A., Heinrich, J., Kreyling, W., and Heyder, J. (2000). Daily mortality and fine and ultrafine particles in erfurt, germany, part i: roleof particle number and particle mass. Res Rep HealthEffInst, 98.
[14] Wood, S. and Augustin, N. (2002). Gams with integrated model selection using penalized regression splines and applications to environmental modelling. Elsevier Science.
[15] Wood, S. A. (2006). Generalized Additive Models:an introduction with R. Chapman and Hall/CRC, London.
[16] Wood, S. A., Goude, S., and Shaw, S. (2014). Supporting material for: Generalized additive models for large datasets. Journal of the Royal Statistical Society.

Author Profile

A.O. Ochugbojuis a M.Sc. degree holder in Statistics and Data Management from University of Essex, UK in 2015 and B.Sc. degree in Mathematics from University of Jos, Nigeria in 2012, and a Lecturer, Department of Mathematics, Federal University Lafia, Nigeria.

A.Yaweholds a Master of Science degree in Statistics and Operations Research from University of Essex, UK in 2015 and Bachelor of Technology degree in Mathematics from ModibboAdama University of Technology Yola, Nigeria in 2005. Currently, a Lecturer, Department of Mathematics and Statistics, Federal University Wukari, Nigeria.

H. A. Odiniya Received the M.Sc. and B.Sc degrees in Operation Research from ModibboAdama University of Technology Yola in 2017 and 2011 respectively and is currently working as a System Analyst at Information and Communication Technology Department Federal University Wukari, Nigeria.

A.A. Musa has Master of Science degree in Statistics and Economics from University of Essex, UK in 2015 and B.Sc. degree in Economics from Ahmadu Bello University Zaria, Nigeria in 2009. Currently works asa Lecturer, Department of Economics, Umar Suleiman College of Education Gashua,Yobe State, Nigeria.