Chapter 5
Big Data and Further Analysis

'Big Data' is a relative term used to describe a tremendously large data. The large data is inclusive of audio, video, unstructured text, social media information, and so much more. Its concept has gained wide publicity or attention in many disciplines. Interestingly, 'Big data' means different things to various disciplines. For example, in atmospheric study, 'big data' means volume of data as large as one terabytes and above. Meanwhile in particle physics, 'big data' is in petabytes and above. For communication outfit, 'big data' may mean zettabytes. Hence, there is the need for disciplinary and multi-disciplinary outfit or research institutes to embrace 'big data' technologies such as in-memory technologies, sensory (Internet of Things) equipment, Cloud Data Storage, magnetic storage, Big Data databases (e.g. MongoDB) etc.

The origin of 'big data' was traced to John Graunt (1663)—a statistical analyst who worked on a large volume of information on bubonic plague in Europe (Foote 2017). Herman Hollerith in 1881 worked with big data in U.S. Census Bureau to create the first big data analysis instrument—Hollerith Tabulating Machine. In other words, the concept of 'big data' has been used in form of machines, processes or applications in the past. In 2005, Roger Mougalas came-up with the word 'big data'. Since its identification and universal acceptability, universities, businesses and governments alike began to establish big data projects. Among application of 'big data', but not limited to the following are: in understanding and targeting consumers; self-optimization; improvement of healthcare; security and law enforcement improvement (Cleverism 2018); the music industry replaces intuition with Big Data studies (Dataversity 2018); being used by cybersecurity to stop cybercrime; to explore the universe using satellite exploration; establishing detailed research outcome such as association rule learning, classification tree analysis, genetic algorithms, machine learning, regression analysis, statistical analysis and social network analysis (Stephenson 2013).

The analytics techniques or methods for understanding big data are many according to specializations and disciplines. Some of analytics techniques that provide the most value for analyzing big data are visualization models, logistic regression, text analytics e.g. information extraction, audio analytics–speech analytics, video content

© Springer Nature Switzerland AG 2019

M. E. Emetere, *Environmental Modeling Using Satellite Imaging and Dataset Re-processing*, Studies in Big Data 54,
https://doi.org/10.1007/978-3-030-13405-1_5

analysis (VCA). Businesses use the visualization models such as Domo, Qlik, Tableau, Sisense, Reltio etc. However, there are some challenges associated with the big data concept. Borne (2014) identified the challenges of big data in an article titled "*top ten big data challenges—a serious look at 10 big data V's*" as: volume of data; variety of data; velocity of data within systems; veracity of data to ascertain its sufficiency to solve problems; validity of the source of data; value of data to specific disciplines; variability of data dynamism; venue showing the destination of data; vocabulary to classify data; and vagueness of data not qualified to be termed 'big data'. Qubole (2008) listed the challenges of big data (facing professionals in different fields of endeavor) as: scalability, lack of talent, actionable insights, data quality, security of data and cost management.

In the next section, the analytics of the large data obtained from MISR for one hundred and thirty locations over the region discussed in the previous chapter was discussed. The visualization model or Computer technique that was used was based on the CERN Root C++ open source.

5.1 Description of Data Source

The dataset was obtained from the Multi-angle Imaging SpectroRadiometer (MISR). As mentioned in chapter one, MISR operates at various directions, that is, nine different angles (70.5°, 60°, 45.6°, 26.1°, 0°, 26.1°, 45.6°, 60°, 20.5°) and gathers data in four different spectral bands (blue, green, red, and near-infrared) of the solar spectrum. The blue band is at wavelength 443 nm, the green band is at wavelength 555 nm, the red band wavelength 670 nm and the infrared band is at wavelength 865 nm. The blue band is used to analyse ice, snow, soil or water. The green band is to analyse Bathymetric mapping and estimating peak vegetation. The red band analyses the variable vegetation slopes and the infrared band analyses the biomass content and shorelines.

Thirteen years dataset was obtained for each of the one hundred and thirty locations across West Africa (see Fig. 5.1). The locations are Benin (Bohicon), Benin (Cotonou), Benin (Kandi), Benin (Natitingou), Benin (Parakou), Benin (Portnovo), Benin (Save), Burkina Faso (Banfora),Burkina Faso (Bobodiolasso), Burkina Faso (Kongoussi), Burkina Faso (Ouagadougou), Burkina Faso (Ouahigouya), Burkina Faso (Dori), Cameroun (Bamenda), Cameroun (Douala), Cameroun (Ebolowa), Cameroun (Garoua), Cameroun (Kousseri), Cameroun (Kumbo), Cameroun (Ngoundere), Cameroun (Younde), Capeverde (Assomada), Capeverde (Ponta), Capeverde (Praia), Chad (Abeche), Chad (Faya), Chad (Mao),Chad (Moundou), Chad (Ndjamena), Chad (Sahr), Cote d'Ivoire (Bondougou), Cote d'Ivoire (Sanpedro), Cote d'Ivoire (Daloa), Cote d'Ivoire (korhogo), Equitorial guinea (Bata), Equitorial guinea (Ebebiyin), Equitorial guinea (Malabo), Gambia (Basse), Gambia (Brikama), Gambia (Farafenni), Gambia (Serekunda), Guinea Bussau (Bafata), Guinea Bussau (Bussau), Guinea Bussau (Gabu), Ghana (Accra), Ghana (Bawku), Ghana (Bolgatanga), Ghana (Sunyani), Ghana (Takoradi), Ghana (Tamale), Guinea

Fig. 5.1 Study area: West Africa and its environ

(Conakry), Guinea (Koundara), Guinea (Macenta), Guinea (Nzerekore), Guinea (Siguiri), Guinea (kankan), Liberia (Buchanan), Liberia (Harper), Liberia (Monronvia), Liberia (Voinjama), Liberia (Yekepa), Mali (Mopti), Mali (Bamako), Mali (Gao), Mali (Kidal), Mali (Nioro), Mali (Segou), Mali (Sikasso), Mauritania (Aqjawajat), Mauritania (Kifah), Mauritania (Nawadibu), Mauritania (Nawaksut), Mauritania (Silibabi), Mauritania (Walatah), Mauritania (Zuwarat), Niger (Agadez), Niger (Arlit), Niger (Gaya), Niger (Magaria), Niger (Niamey), Niger (Tahoua), Nigeria (Abeokuta), Nigeria (Abuja), Nigeria (Calabar), Nigeria (Damaturu), Nigeria (Enugu), Nigeria (Gusau), Nigeria (Ibadan), Nigeria (Ikot Ekpene), Nigeria (Ilorin), Nigeria (Jos), Nigeria (Kaduna), Nigeria (Kano), Nigeria (Kastina), Nigeria (Lagos), Nigeria (Minna), Nigeria (Mubi), Nigeria (Ogbomoso), Nigeria (Ondo), Nigeria (Oshogbo), Nigeria (Sokoto), Nigeria (Onitsha), Nigeria (Owerri), Nigeria (Warri), Senegal (Dakar), Senegal (Louga), Senegal (Tambakounda), Senegal (Ziguinchor), Sierra Leone (Makeni), Sierra Leone (Binkolo), Sierra Leone (Kabala), Sierra Leone (Talama), Togo (Atakpame), Togo (Dapaong), Togo (Kara), Togo (Lome) and Togo (Sokode).

5.2 Data Analysis: Relevant Connection to Imagery

In this section, the discussion is based on specific location of interest as depicted in Chap. 4. The format of the plot is as follows: the first subsection tagged 'a' show the 3D image of AOD at 550 nm against the mean AOD (of the blue, green, red

and infra-red band); the second subsection tagged 'b' show the 3D image of AOD 865 nm against Mean AOD; the third subsection tagged 'c' show the 2D image of AOD 440 nm against sum AOD (550, 670, 865 nm); the fourth subsection tagged 'd' show the 2D image of AOD 550 nm against sum AOD (440, 670, 865 nm); the fifth subsection tagged 'e' show the 2D image of AOD 670 nm against sum AOD (440, 550, 865 nm); the sixth subsection tagged 'f' show the 2D image of AOD 865 nm against sum AOD (440, 550, 670 nm); the seventh subsection tagged 'g' show the scattered plot of AOD at 440 nm against number of days; the eighth subsection tagged 'h' show the scattered plot of AOD at 550 nm against number of days; the ninth subsection tagged 'i' show the scattered plot of AOD at 670 nm against number of days; the tenth subsection tagged 'j' show the scattered plot of AOD at 865 nm against number of days; the eleventh subsection tagged 'k' show the 3D image of AOD at 865 nm against AOD at 670 nm; the twelfth subsection tagged 'l' show the 3D image of AOD at 550 nm against AOD at 440 nm; the thirteenth subsection tagged 'm' show the 3D image of AOD at 670 nm against AOD at 550 nm; the fourteenth subsection tagged 'n' show the 3D image of AOD at 865 nm against AOD at 440 nm; the fifteenth subsection tagged 'o' show the 3D image of AOD at 670 nm against AOD at 440 nm; the sixteenth subsection tagged 'p' show the 3D image of AOD at 865 nm against AOD at 670 nm; the seventeenth subsection tagged 'q' show the 2D image of AOD at 440, 550, 670 and 865 nm against the number of days between 2000 and 2013.

In Fig. 4.43, there was high aerosol concentration in some parts on Benin. To understand the extent of pollution, Parakou and Save were chosen as shown in Figs. 5.2 and 5.3 respectively. Figure 5.2a reveals a linear connection between the AOD at 550 with the mean AOD. This may possibly mean that the deviation from the mean is insignificant or very low. Hence, in reality, much assertion cannot be made on the possibility of comparing the AOD at 550 nm and mean AOD to understand the individual AOD expressed in Eq. (1.7). Figure 5.2b show significant scattered distribution. Hence, AOD at 865 may be use to understand the AOD of individual aerosol components expressed in Eq. (1.7). While the AOD at 440 nm against the sum of other band-AOD showed scattered distribution (Fig. 5.2c), the AOD at 550 nm against the sum of other band-AOD showed a linear representation (Fig. 5.2d). Figure 5.2e show the linear relationship between AOD at 670 nm and sum of AOD that represents other AOD band. AOD at 865 nm (like AOD at 440 nm) had a scattered distribution as observed in Fig. 5.2f. The scattered plot for all band is shown in Fig. 5.2g–j. AOD at 865 nm had the most coherent distribution (see color map representation in Fig. 5.2j). AOD at 670 nm (Fig. 5.2i), AOD at 440 nm (Fig. 5.2g) and AOD at 550 nm (Fig. 5.2h) had coherent distribution in descending order.

The interdependency of the AOD of each band is discussed in Fig. 5.2k–p. The interpretation of Fig. 5.2k–p is done from the surface and the shapes within the 3D image. The plot of AOD at 865 nm and AOD at 670 nm (Fig. 5.2k) has high AOD presence between 0.1 and 0.6. The highest and lowest AOD frequency can be found in 0.2–0.3 and 0.8–1.2 respectively. This means the combination of AOD at 865 and AOD at 670 will yield low results when used to estimate aerosol parameters e.g. Angstrom exponent. The plot of AOD at 550 nm and AOD at 440 nm (Fig. 5.2l) have

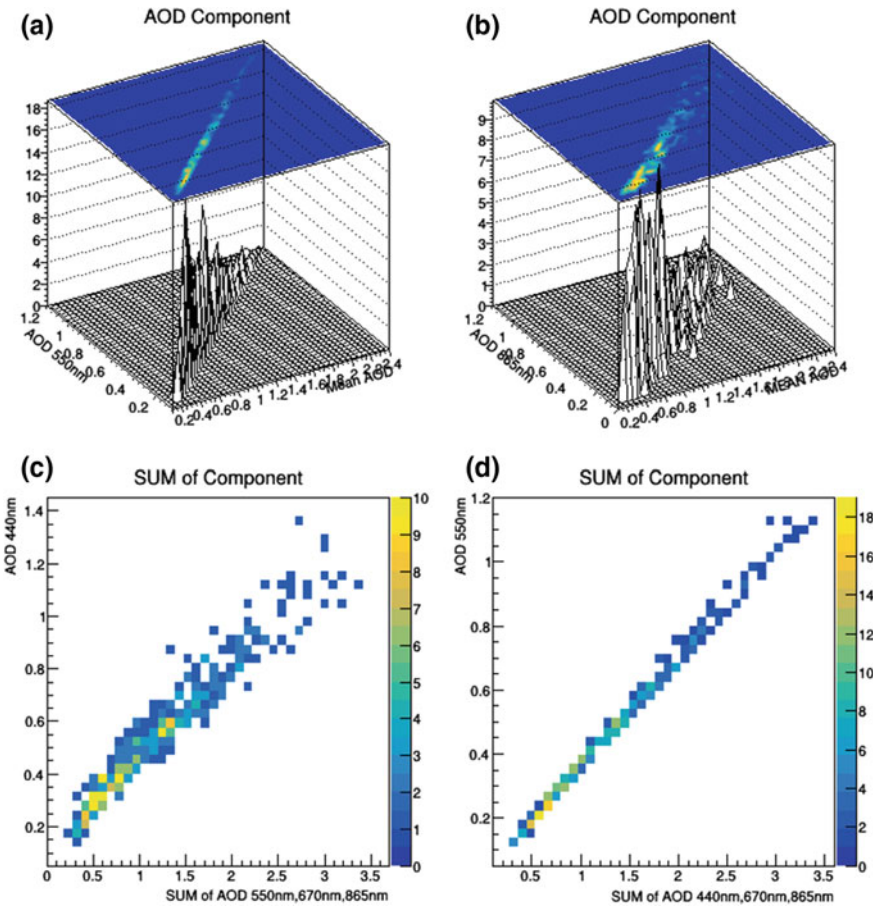

Fig. 5.2 **a**–**d** Aerosol inter-relationality I. **e**–**h** Aerosol daily performance I. **i**–**l** Aerosol inter-relationality II. **m**–**p** Aerosol inter-relationality III. **q** Virtual performance of individual AOD

high AOD presence between 0.1 and 0.85. The highest and lowest AOD frequency can be found in 0.2–0.8 and 1.0–1.2 respectively. This means that the Angstrom exponent for each point will be more.

The plot of AOD at 670 nm and AOD at 550 nm (Fig. 5.2m) have high AOD presence between 0.2 and 0.4. The highest and lowest AOD frequency can be found in 0.2–0.5 and 1.0–1.2 respectively.

This means that the Angstrom exponent for each point will be less. Also, Fig. 5.2m is adjudged the most linear and scanty plot. Hence, AOD at 670 and 550 nm has the less dependency on each other. The plot of AOD at 865 nm and AOD at 440 nm (Fig. 5.2n) have high AOD presence between 0.2 and 0.9. The highest and lowest AOD frequency can be found in 0.2–0.9 and 1.0–1.2 respectively. The uniqueness

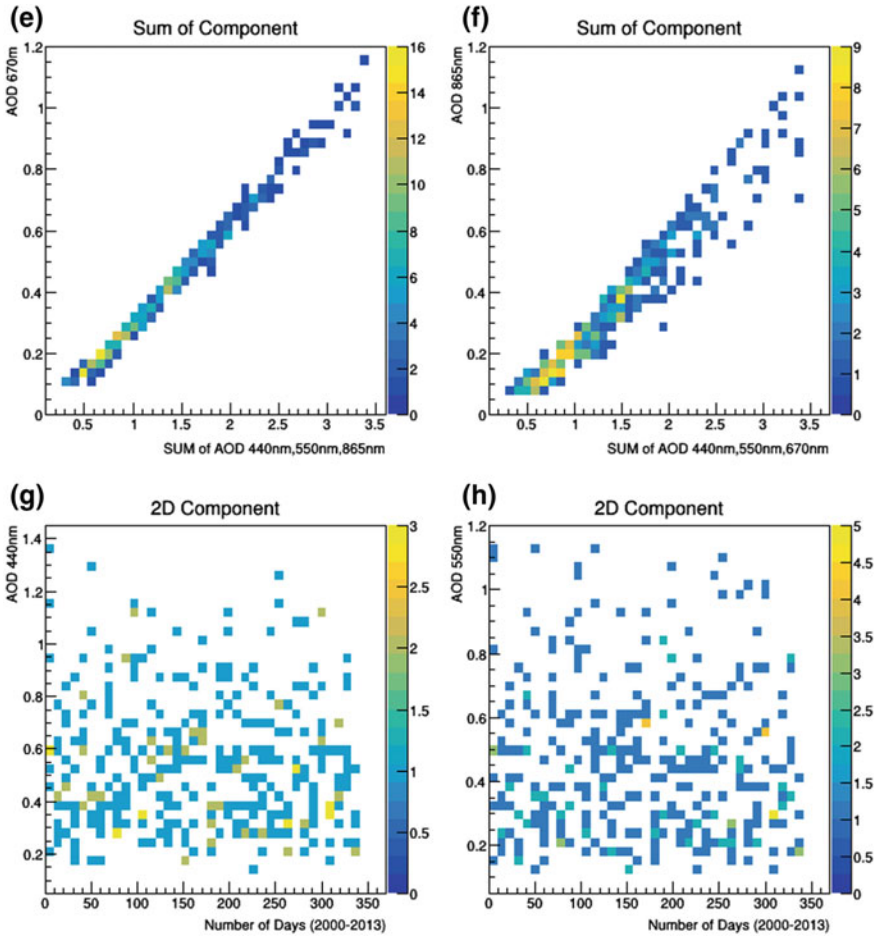

Fig. 5.2 (continued)

of this plot lies in the scattered distribution of its components. Unlike Fig. 5.2m, n has the highest dependency on each other.

The plot of AOD at 670 nm and AOD at 440 nm (Fig. 5.2o) has high AOD presence between 0.2 and 0.7. The highest and lowest AOD frequency can be found in 0.2–0.7 and 1.0–1.2 respectively. Figure 5.2p also shares almost same traits as Fig. 5.2o. The difference between both diagrams is that the scattering of the AOD points in Fig. 5.2p spreads out of its linearity. On the other hand, Fig. 5.2o shows higher dependency on the AOD bands i.e. next to Fig. 5.2n. Figure 5.2q shows the individual performance of the AOD band between 2000 and 2013. Days that had AOD ≥ 1.0 was documented to understand the sequence at which the anthropogenic source influences the AOD's magnitude.

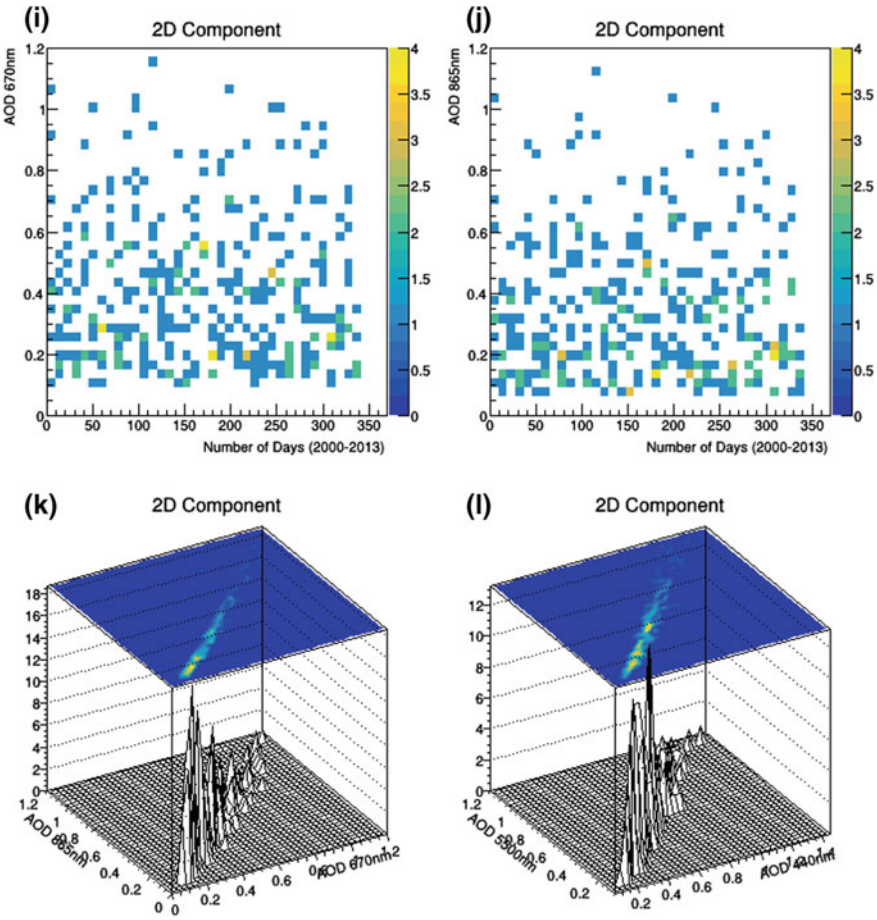

Fig. 5.2 (continued)

The peak of ≥1.0 (of AOD) were noted on the following days: 16-March-2000 (day-3), 22-March-2002 (day-50), 9-March-2003 (day-71), 11-March-2004 (day-96), 18-March-2004 (day-97), 5-March-2005 (day-121), 10-March-2006 (day-147), 4-March-2007 (day-175), 15-March-2008 (day-201), 3-March-2010 (250), 21-March-2010 (day-252), 11-Feb-2011 (day-273), 7-April-2011 (day-280), 5-Feb-2012 (day-299), and 17-March-2012 (day-304). It could be concluded that the highest AOD over Parakou occurs every March of the year. No peaks appeared in 2009 and 2013 while two peaks appeared in 2011 within two unusual months (i.e. February and April). It can be inferred that 2009 is unique in the sense that the anthropogenic pollution or aerosol retention might have decreased.

Figure 5.3a–f has same technical concept as Fig. 5.2a–f. The scattered plot for all band is shown in Fig. 5.3g–j differ from Fig. 5.2g–j. Figure 5.3h had the most

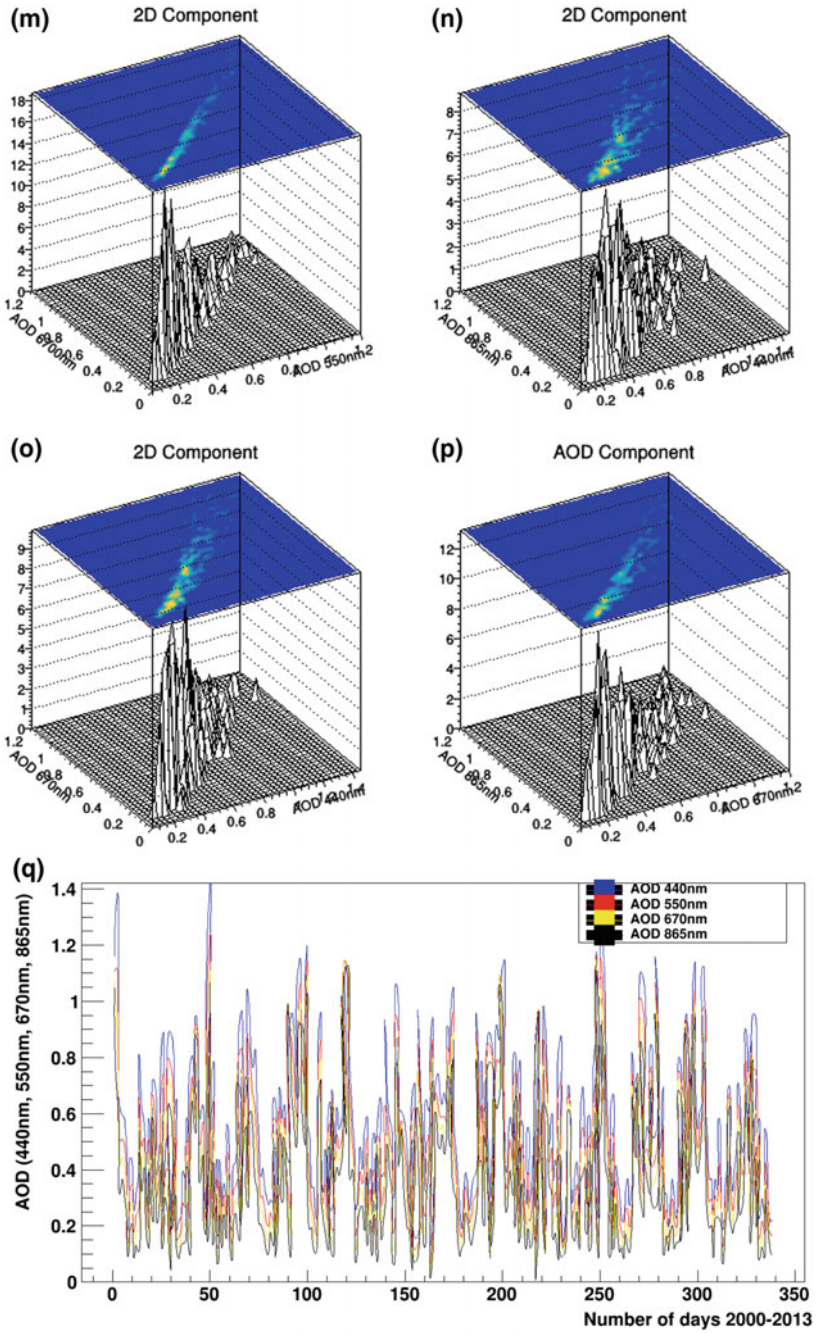

Fig. 5.2 (continued)

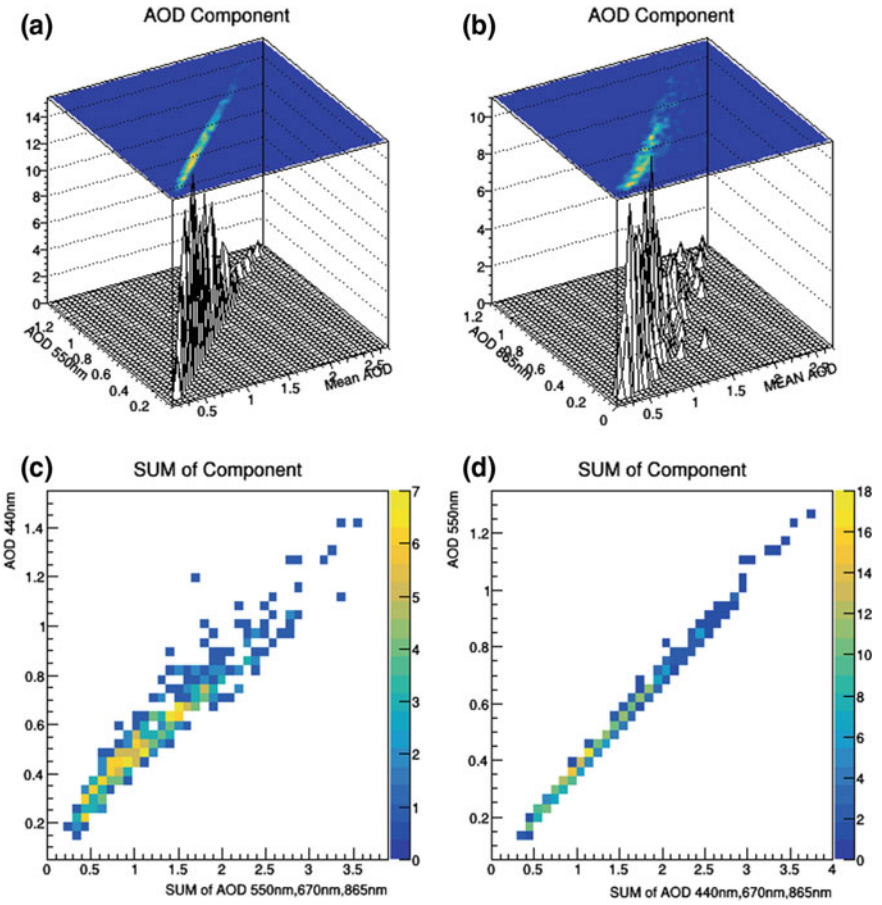

Fig. 5.3 **a–d** Aerosol inter-relationality I. **e–h** Aerosol daily performance I. **i–l** Aerosol inter-relationality II. **m–p** Aerosol inter-relationality III. **q** Virtual performance of individual AOD

coherent distribution (see color map representation in Fig. 5.3h). AOD at 875 nm (Fig. 5.2j), AOD at 440 nm (Fig. 5.2g) and AOD at 670 nm (Fig. 5.2i) had coherent distribution in descending order. The interdependency of the AOD (Fig. 5.3k–p) of each band has same trend as Fig. 5.2k–p.

The peaks of $\geq$1.0 (of AOD) were noted on the following days: 7-March-2000 (day-2), 21-Jan-2001 (day-20), 11-April-2001 (day-28), 24-Jan-2002 (43), 22-March-2002 (day-51), 16-March-2003 (day-74), 11-March-2004 (day-96), 29-Nov-2004 (day-107), 17-Feb-2005 (day-109), 5-March-2005 (day-121), 19-Jan-2006 (day-139), 20-Feb-2006 (day-144), 4-March-2007 (day-174), 26-Dec-2008 (day-219), 4-Feb-2011 (day-273), 8-March-2011 (day-275), 19-Dec-2011 (day-287), 20-Jan-2012 (day-293), 17-March-2012 (day-300).

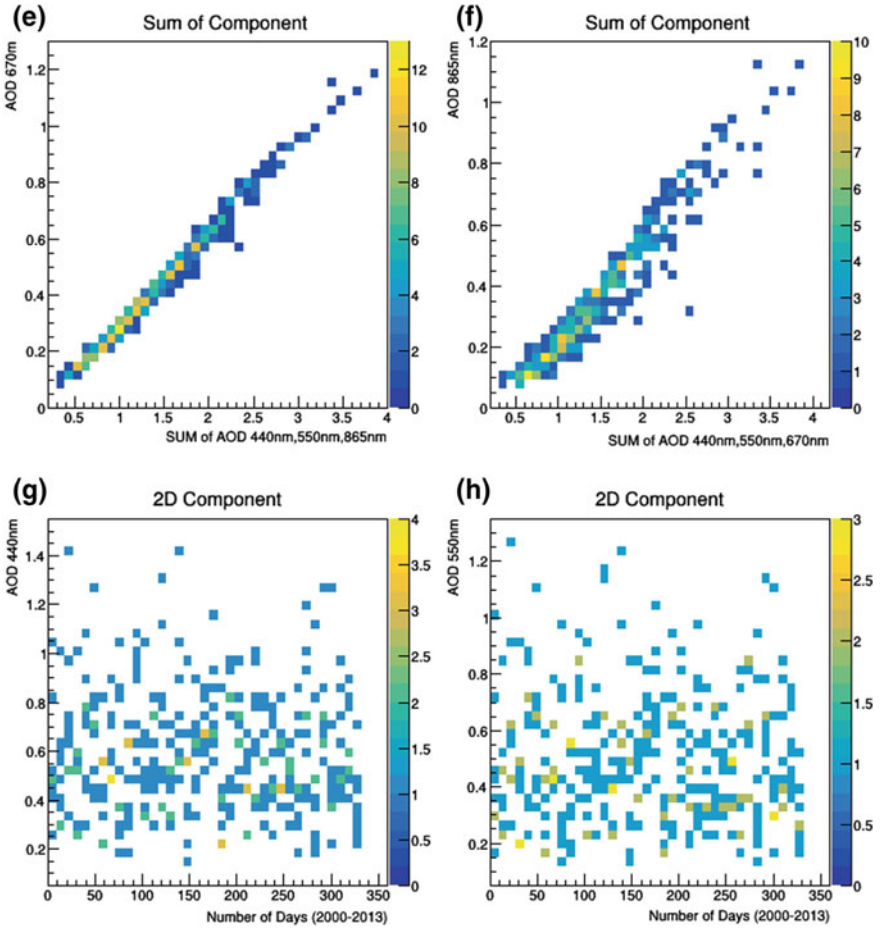

Fig. 5.3 (continued)

It could be concluded that the highest AOD over Save occurs in the month of March over the years. However, there were no peaks in 2009, 2010 and 2013 while two peaks appeared in 2001, 2002, 2004, 2005, 2011 and 2012. It is therefore corroborated by Fig. 5.3q that 2009 had decreased anthropogenic pollution or aerosol retention.

Figures 4.8, 4.15, 4.29 and 4.43 constantly showed the influence of the Sahara Desert at the boundaries of Niger and Chad. It is clear that aside Sahara dust, there are human activities that led to the deposition of sulfate and black carbon over the region. It is in this light we considered close town in Chad (such as Faya and Mao) and Niger (Agadez).

Figure 5.4a–f has same technical concept as Fig. 5.2a–f. The scattered plot for all band is shown in Fig. 5.4g–j. Compared to Figs. 5.2g–j and 5.3g–j, the scattered plot in Fig. 5.4g–j has more data points because its satellite dataset retrieval is not hindered

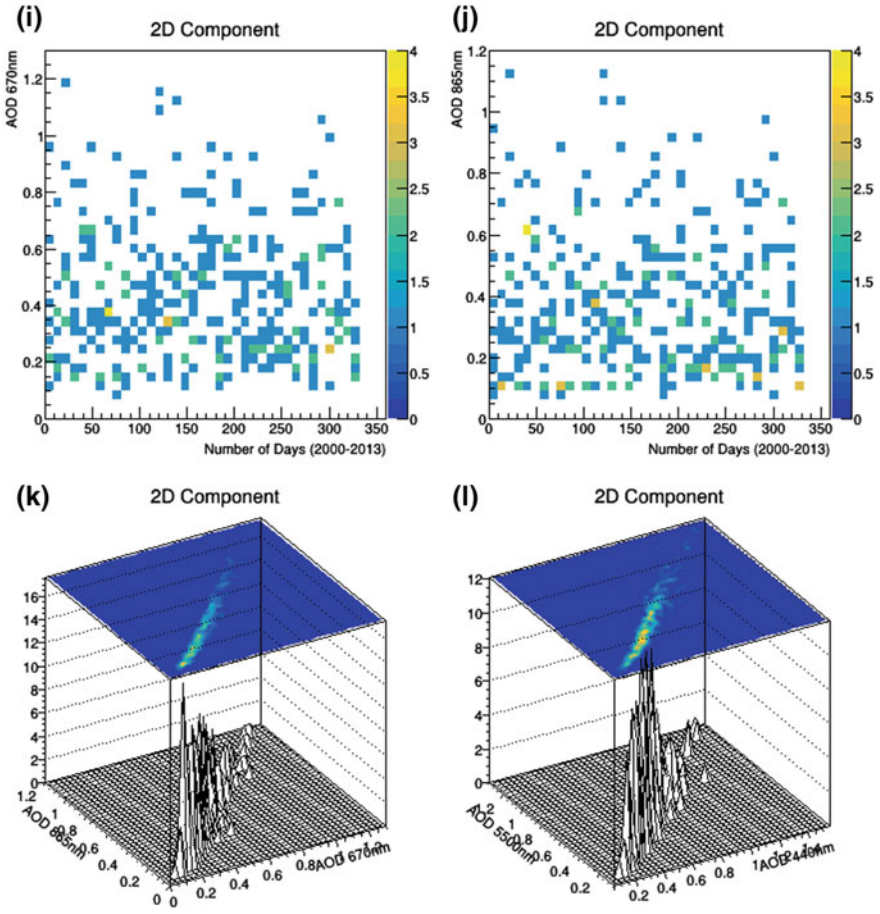

Fig. 5.3 (continued)

by moist (Emetere 2016a, b). Satellite exploration do not retrieve dataset for everyday of the year because of technical issues as orbiting time, moist etc. This occurrence is referred to as 'data loss'. It has been observed that there is a large volume of 'data loss' over West Africa (Emetere et al. 2015a, b, c, d, 2016; Emetere and Akinyemi 2017). Also, there are more 'data loss' in southern-coastline parts of West Africa than in its northern part. Figure 5.4i had the most coherent distribution (AOD at 670 nm). AOD at 550 nm (Fig. 5.4h), AOD at 865 nm (Fig. 5.4j) and AOD at 440 nm (Fig. 5.4g) had coherent distribution in descending order. The interdependency of the AOD (Fig. 5.4k–p) of each band has low coincidence with one another. This means the generation of the dataset for each bands are very dynamic due to the large dispersion source (Sahara Desert).

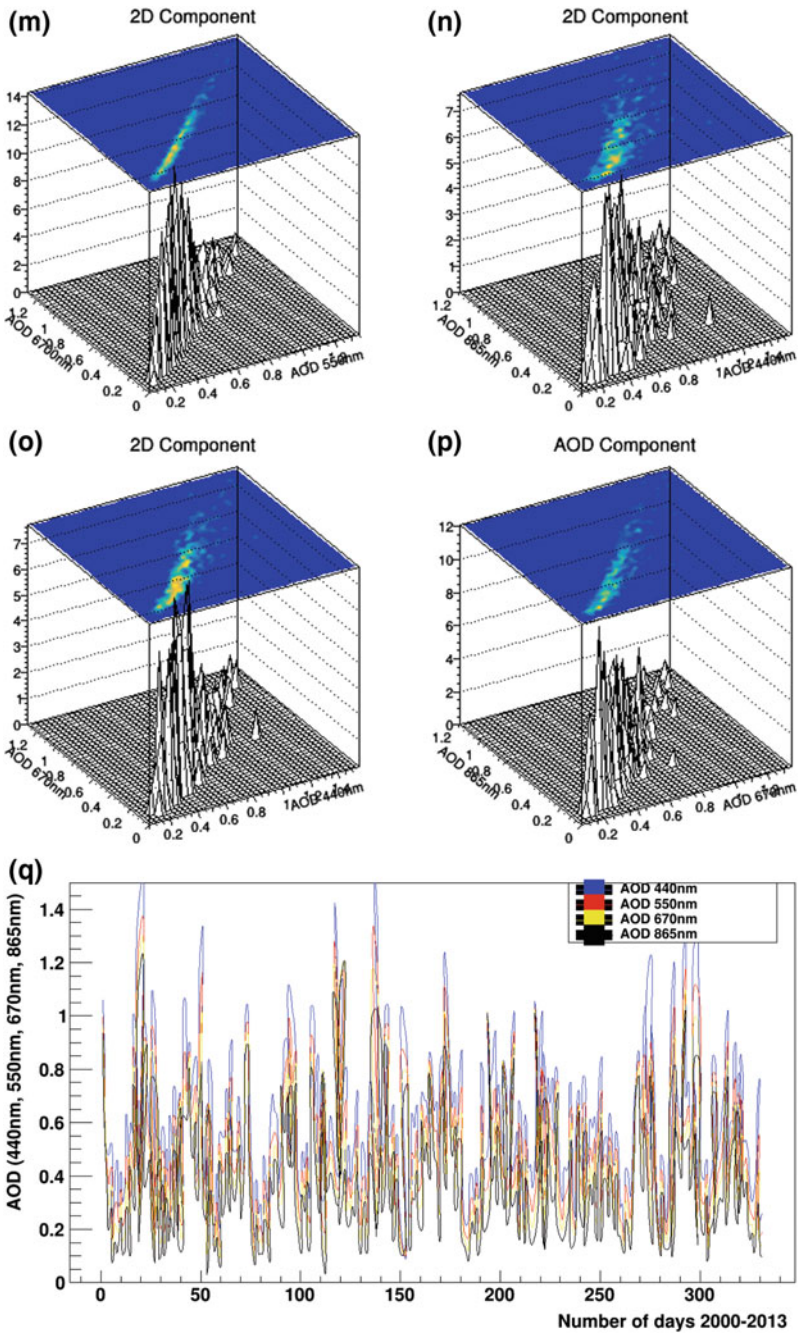

Fig. 5.3 (continued)

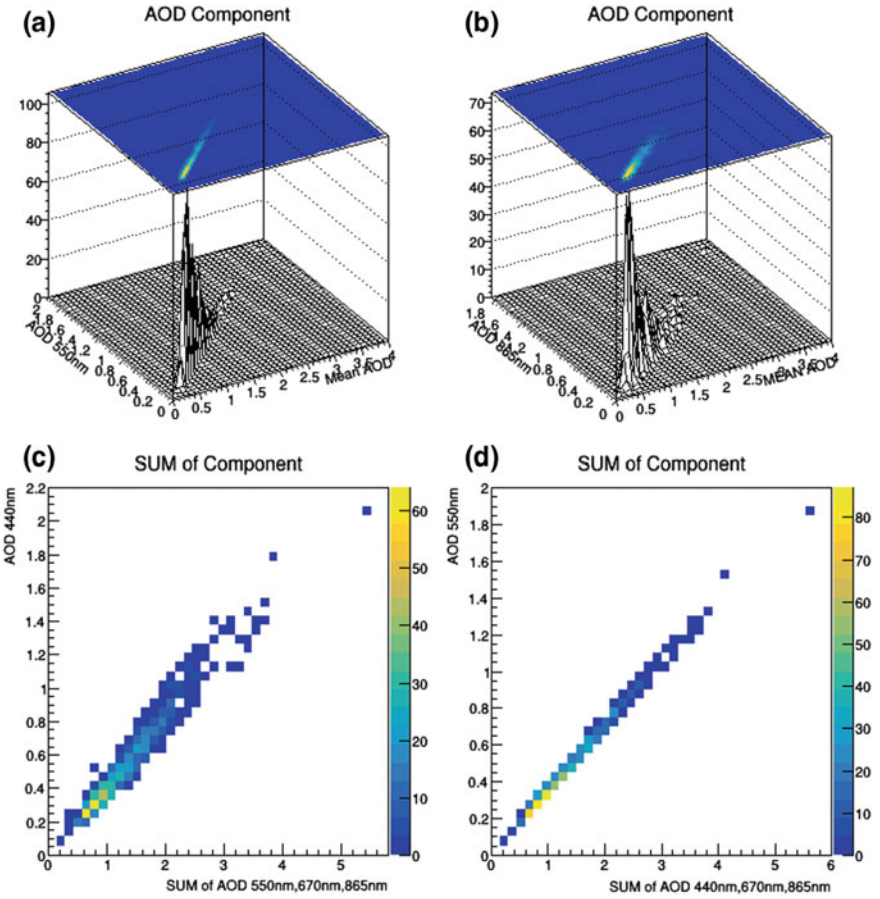

Fig. 5.4 **a**–**d** Aerosol inter-relationality I. **e**–**h** Aerosol daily performance I. **i**–**l** Aerosol inter-relationality II. **m**–**p** Aerosol inter-relationality III. **q** Virtual performance of individual AOD

Figure 5.4q has more unique feature i.e. sinusoidal. This result is same for Moa-Chad (Appendix: Fig. B.1) and Agadez-Niger (Appendix: Fig. B.2). The comparative analysis of the three locations (Faya, Mao and Agadez is shown in Table 5.1. The date presented are days when the AOD over the location exceeds 1 i.e. AOD $\geq$ 1. As discussed earlier, there were lots of data loss, so the number of day presented (for example day-24) refer to the day the satellite retrieved meaningful dataset during the year. The observation drawn from Table 5.1 gives detail on the satellite imagery shown in Figs. 4.8, 4.15, 4.29 and 4.43.

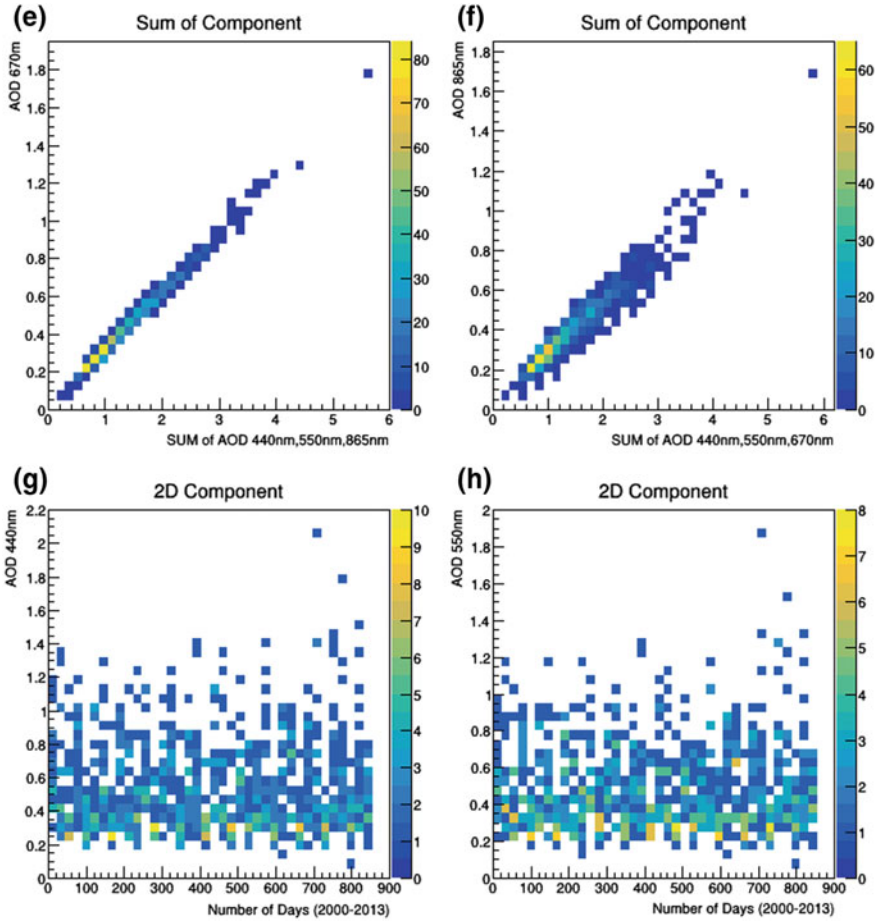

Fig. 5.4 (continued)

The observations are:

i. Satellite AOD retrieval takes place in the following order i.e. Faya-Chad, Mao-Chad and Agadez-Niger. This may be adduced to the satellite orbiting time over the locations;
ii. Mao-Chad had the most frequency of AOD ≥ 1;
iii. Faya-Chad had more dataset than Mao-Chad and Agadez-Niger;
iv. In 2003, 2005 and 2007, all the location had AOD ≥ 1 in the month of April;
v. Satellite AOD retrieval takes place in the reversed order i.e. Agadez-Niger, Faya-Chad and Mao-Chad. This may be adduced to the satellite orbiting time over the locations or aerosol layer transport in the atmosphere;

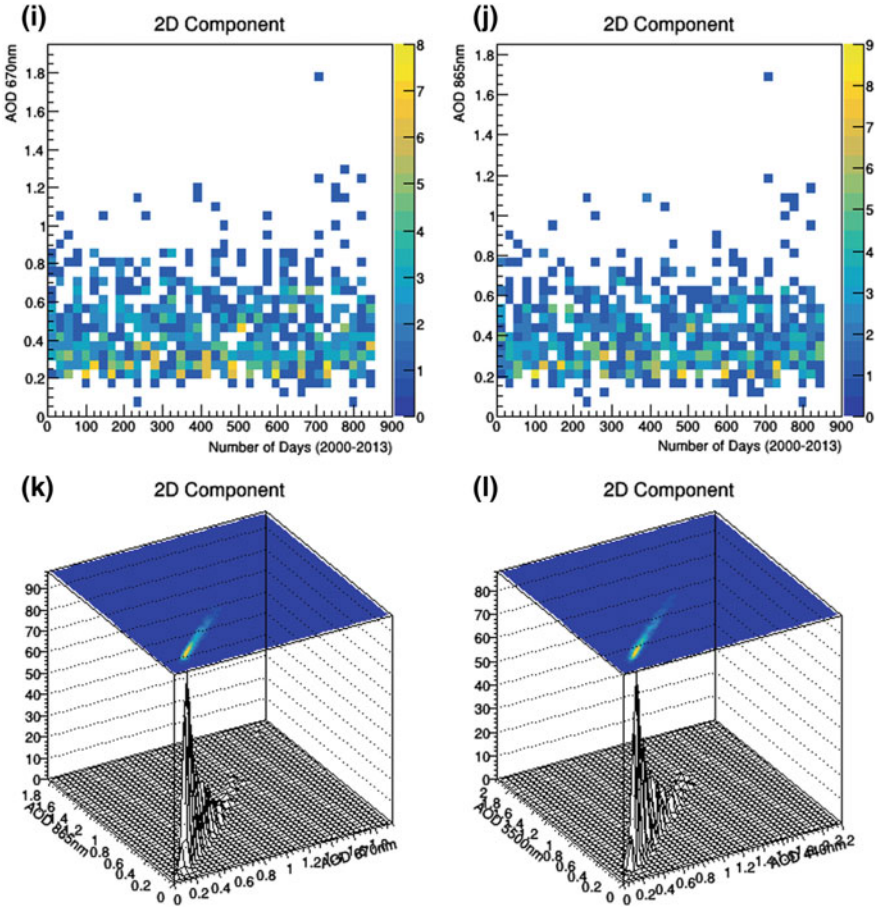

Fig. 5.4 (continued)

vi. The most active month in the three locations is April. The active months are listed in descending order i.e. June, August and July. Emetere (2016a, b, 2017a, b, c, d) postulated that the months mentioned may be due to aerosol retention from October to March.

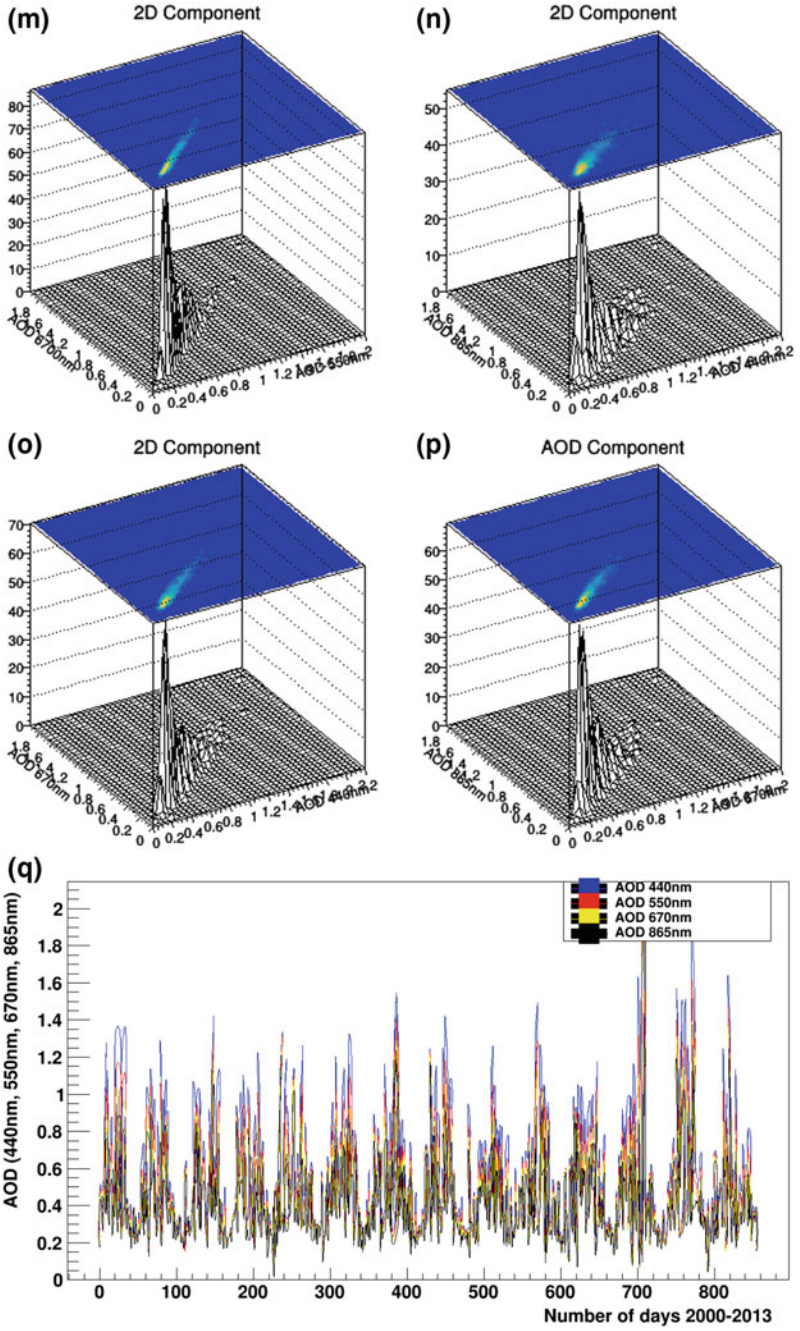

Fig. 5.4 (continued)

5.3 Data Analysis: General Comments on Locations

The interpretation of simulation in environmental science (as shown above) depends on the minor details within its confines and the prevailing situation that may generally lead to a minute or significant change in the simulation. For example, the specific location that were selected in chapter five was based on the information that was retrieved from the satellite images in chapter four. The quality of interpretation is directly proportional to the information that can be retrieved from the available simulations.

One of the cardinal focus of this book is to maximum information that can be retrieved from 'big data'. For example, if a terabyte datasets of images (say >30,000) is to be analyzed, it is advisable to use the following rules:

i. Do the statistical analysis of the images (see Figs. 4.53–4.55). This process engenders more information on the general trends of the curves. The trend may either satisfy a given condition or known event. For example, aerosols optical depth is expected to be high from October to April in the tropical region of west Africa. Also, the sudden increase in aerosol retention from June to August (Fig. 5.1) gives indication of the existence of an unknown event in the geographical area. Hence, seeking adequate interpretation for the trends may allow you narrow to specific points in the simulation.
ii. Do the multi-dimensional analysis of each images located on the hotspot. For example, Figs. 4.1–4.52 will show the significant differences or similarities between chosen images. This section can only be successful if adequate attention is given to the program or macro that is used for the simulation. It is advisable to numerically analyze the images so you can quantify the 'big data' inform of computational or mathematical model. For example, Figs. 5.5 and 5.6 show the intensity and deviations obtained from 3600 images. Many more processes that can lead to more quality information can be extracted. This information can be categorized as derived and basic. Figure 5.5 is a basic information that can be obtained directly from all the images while Fig. 5.6 is the product of a basic parameter (i.e. derived information). Figure 5.6 may be obtained from the characterization of the pixel of each images or the intensity of the images. The author, believes that the deviation via pixel characterization is the best way of spotting the deviation between images. An example of the first fifteen pixel of three grayscale images is shown Figs. 5.7.
iii. The validation of the results obtained from the re-processed satellite images was achieved by considering the analysis of the '.txt' file for each location. Aside validating assertions, it can be used to obtain detailed information of the hotspot. For example, Table 5.1 show that more anthropogenic activity occurs in Chad (Mao) within the three locations that is close to the Sahara Desert. When designing a project in environmental science (may be using 'big data'), it is advisable to provide means of validation. Also, it is necessary to design the program or macro to generate formidable simulations that will ultimate improve the quality of the project.

Table 5.1 Interdependency of cities close to the Sahara at AOD ≥ 1

Year	Chad (Faya)	Chad (Mao)	Niger (Agadez)
2000	20-April-2000 (day-9), 1-August-2000 (day-25)	6-August-2000 (day-20), 25-September-2000 (day-28)	Nil
2001	21-April-2001 (day-68)	19-April-2001 (day-56), 30-May-2001 (day-60)	2-June-2001 (day-46)
2002	26-April-2002 (day-129), 23-August-2002 (day-148)	21-March-2002 (day-103), 1-May-2002 (day-107), 13-July-2002 (day-118), 6-September-2002 (day-125), 22-September-2002 (day-126)	12-June-2002 (day-87)
2003	20-April-2003 (day-189), 10-August-2003 (day-208)	4-February-2003 (day-148), 11-April-2003 (day-159), 13-May-2003 (day-164), 27-May-2003 (day-166)	28-April-2003 (day-119)
2004	5-March-2004 (day-237), 31-May-2004 (day-253), 27-July-2004 (day-264)	16-February-2004 (day-201), 13-May-2004 (day-213), 16-July-2004 (day 221)	1-June-2004 (day-157)
2005	25-April-2005 (day-309), 8-August-2005 (day-328)	23-April-2005 (day-260)	26-April-2005 (day-193)
2006	28-April-2006 (day-372), 24-July-2006 (day-388)	9-March-2006 (day-300), 3-May-2006 (day-309)	16-June-2006 (day-239)
2007	6-April-2007 (day-432), 27-July-2007 (day-452)	13-April-2007 (day-362), 15-May-2007 (day-367), 2-July-2007 (day-376), 6-October-2007 (day-387)	7-April-2007 (day-268)
2008	6-July-2008 (day-511)	26-May-2008 (day-416), 29-September-2008 (day-432)	18-April-2008 (day-306)
2009	23-June-2009 (day-569)	24-March-2009 (day-456), 30-July-2009 (day-471), 8-August-2009 (day-472)	17-July-2009 (day-355)
2010	23-April-2010 (day-624), 14-September-2010 (day-647)	20-March-2010 (day-508)	8-April-2010 (day-383)
2011	7-August-2011 (day-707)	15-April-2011 (day-561), 20-June-2011 (day-570)	26-March-2011 (day-422)
2012	6-June-2012 (day-762), 9-August-2012 (day-773)	16-March-2012 (day-608), 18-March-2012 (day-609)	7-June-2012 (day-472)
2013	8-May-2013 (day-820)	7-June-2013 (day-663), 25-July-2013 (day-668)	3-June-2013 (day-514)

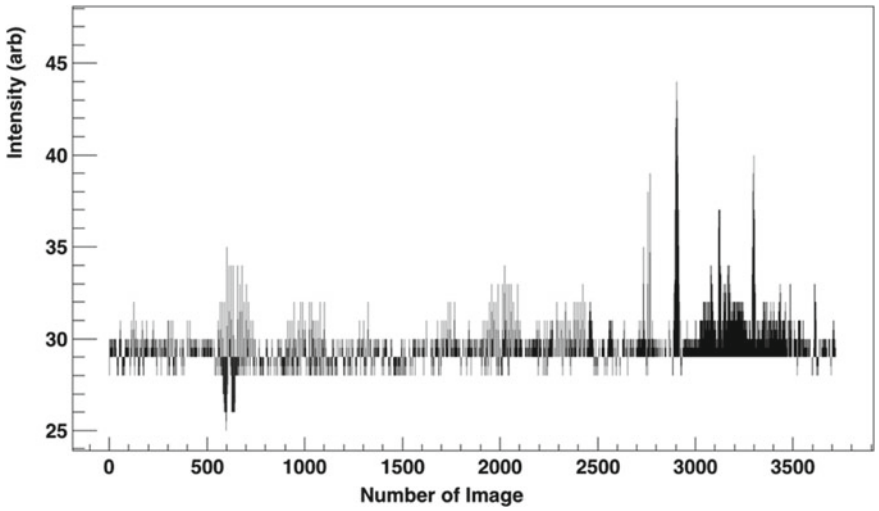

Fig. 5.5 The intensity obtained from 3600 images

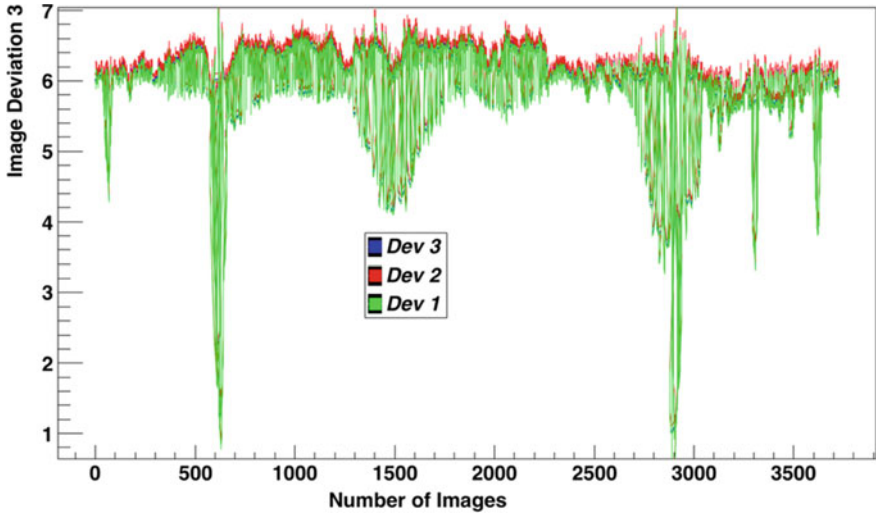

Fig. 5.6 The deviations obtained from 3600 images

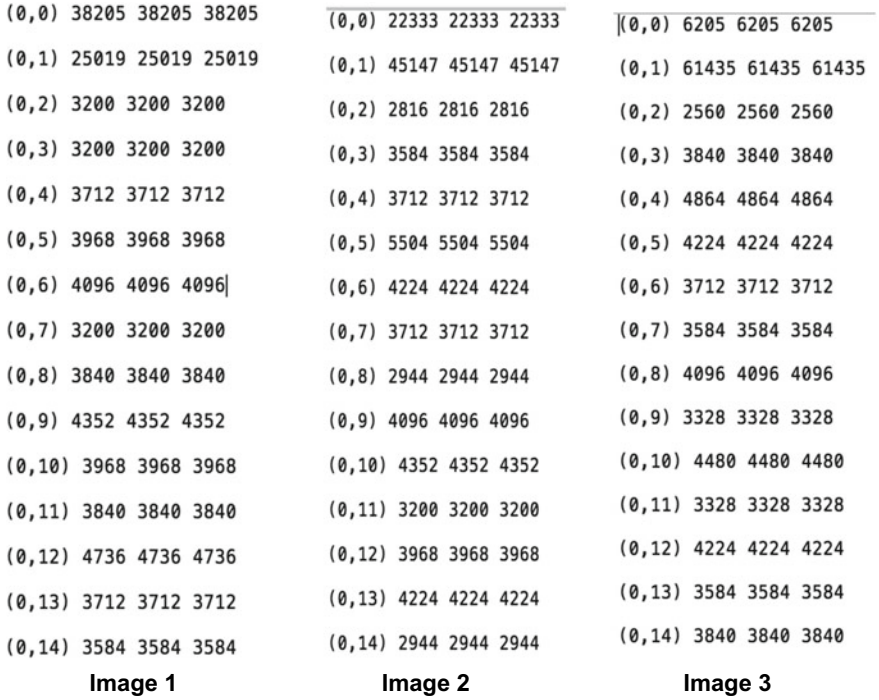

Fig. 5.7 Characterization of image pixel

5.4 Designing the Code to Analyze Big Data

The modalities in this section is same as the Sect. 4.3. In this section, two types of macros shall be described. The format of big data discussed in this section is '.txt', '.dat' etc. The data may come in a raw form i.e. many unwanted information within rows or columns. In actual fact, most datasets obtained from primary sources like satellite companies, communication companies etc. are in a raw format which may be irritating to edit. Editing a raw dataset of one terabyte and above is tedious and fool-hardy. The macro below shows how the author edited raw dataset within the code.

5.4.1 Macro One

```
void Macro1(){
    Float_t Channel,ChannelContent,z,Channel1,ChannelContent1,z1;
    Int_t ncols,ncols1;
    Int_t nlines = 0,nlines1 = 0;
    char line[3000];  // As large as the lines that you are reading in

    // Read from .dat file.

    ifstream datFile;
    ifstream datFile1;
    datFile.open("Space_ValveClosed_mca1.dat");
    datFile1.open("Space_ValveOpen_mca1.dat");
    // skip the first two lines of the dataset


    datFile.getline(line,128);
    datFile.getline(line,128);
    datFile1.getline(line,128);
    datFile1.getline(line,128);

    TFile*f = new TFile("Space_ValveClosed_mca1","RECREATE");
```

```
TFile *f1 = new TFile("Space_ValveOpen_mca1","RECREATE");
TH1F *h = new TH1F("h","Channel distribution",100,-4,4);
TH1F *h1 = new TH1F("h1","Channel distribution",100,-4,4);
TNtuple *ntuple = new TNtuple("ntuple","data from MCA","Channel:ChannelContent");
TNtuple *ntuple1 = new TNtuple("ntuple1","data from MCA","Channel1:ChannelContent1");
while (1) {
    datFile >> Channel >> ChannelContent;
    if (!datFile.good()) break;
    if (nlines < 5) printf("x=%8f, y=%8f\n",Channel,ChannelContent);
    h->Fill(Channel);
    ntuple->Fill(Channel,ChannelContent);
    nlines++;
}
printf(" found %d points\n",nlines);
 datFile.close();

  printf(line, "%f\n",z);

while (2) {
    datFile1 >> Channel1 >> ChannelContent1;
    if (!datFile1.good()) break;
    if (nlines1 < 5) printf("x=%8f, y=%8f\n",Channel1,ChannelContent1);
    h->Fill(Channel1);
    ntuple->Fill(Channel1,ChannelContent1);
    nlines1++;
}
printf(" found %d points\n",nlines1);
 datFile1.close();

  printf(line, "%f\n",z);

      TCanvas *MyCanvas = new TCanvas("canv", "General Plots",800,600);
```

```
        TCanvas *MyCanvas1 = new TCanvas("canv1", "General Plots",800,600);

        ntuple->Draw("ChannelContent:Channel");
        ntuple1->Draw("ChannelContent1:Channel1");

        TCanvas *MyCan = new TCanvas("ca", "General Plots",800,600);
        TCanvas *MyCan1 = new TCanvas("ca1", "General Plots",800,600);
        //MyCanvas->Divide(2,1);
        //MyCanvas->cd(1);
        ntuple1->Draw("ChannelContent/2075.007998:Channel");
        TH2F *htemp = (TH2F*)gPad->GetPrimitive("htemp");
        htemp->GetXaxis()->SetTitle("Channel");
        htemp->GetYaxis()->SetTitle("Channel Content per time [count/sec]");
        htemp->SetFillColor(42);
       htemp->SetMarkerColor(3);
    htemp->SetMarkerStyle(3);
            htemp->SetTitle("MCA Plots");
             f->Write();
}
```

5.4.2 Macro Two

The second type of dataset is the processed dataset. This kind of data can be obtained from environmental monitoring centers etc. Irrelevant information is littered within rows and column. Hence, the way of writing the codes differs as shown below.

```
void Macro2(const char *dirname="testdata.txt", const char *ext=".dat"){
    TString dir = gSystem->UnixPathName(gInterpreter->GetCurrentMacroName());
      dir.ReplaceAll("Macro2.C","");
      dir.ReplaceAll("/./","/");
      TFile *f = new TFile("dirname.root","RECREATE"); //create file data.root
      TTree *tree = new TTree("tree","data from ascii file");
      Double_t nlines = tree-
>ReadFile(Form(dirname,dir.Data()),"pa:pb:pc:pd:pe:pf:pg:ph:pi:pj:pk;pl:pm:pn:po:pp:pq:pr:
```

```
ps:pt:pu:pv:pw:px:py,pz");//create tree with
          gROOT->SetStyle("Plain");
          gStyle->SetOptStat(1111);
          gStyle->SetOptFit(1111);
          TCanvas *c = new TCanvas("c", "General Plots1",800,600);
          c->Divide(2,2);
          c->SetFillColor(5);
          c->SetFrameFillColor(10);
          TMultiGraph * mg = new TMultiGraph("mg","mg");
        //make graphs
          c->cd(1);
         tree->Draw("pb:pa");
          TGraph *gr = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
          gr->SetName("myGraph");
         /*TH2F *htemp = (TH2F*)gPad->GetPrimitive("htemp");
         htemp->GetXaxis()->SetTitle("Channel");
         htemp->GetYaxis()->SetTitle("Channel Content per time");
         htemp->SetFillColor(42);
        htemp->SetMarkerColor(3);
    htemp->SetMarkerStyle(7);
               htemp->SetTitle("MCA Plots");*/
          gr->Draw();

        c->cd(2);
        tree->Draw("pc:pa");
        TGraph *gr2 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
         gr2->SetName("myGraph");
        gr2->Draw();
        c->cd(3);
        tree->Draw("pd:pa");
         TGraph *gr3 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
        gr3->SetName("myGraph");
        gr3->Draw();
```

```
    c->cd(4);
    tree->Draw("pe:pa");
    TGraph *gr4 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
    gr4->SetName("myGraph");
    gr4->Draw();
    c->Update();
  TImage *img = TImage::Create();
   img->FromPad(c);
   img->WriteImage("canvas1.png");


    TCanvas *c1 = new TCanvas("c1", "General Plots2",800,600);
    c1->Divide(2,2);
    c1->cd(1);
    tree->Draw("pf:pa");
    TGraph *jr = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
    jr->SetName("myGraph");
    jr->Draw();

    c1->cd(2);
    tree->Draw("pg:pa");
    TGraph *jr2 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
    jr2->SetName("myGraph");
    jr2->Draw();

    c1->cd(3);
    tree->Draw("ph:pa");
    TGraph *jr3 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
    jr3->SetName("myGraph");
    jr3->Draw();

    c1->cd(4);
    tree->Draw("pi:pa");
    TGraph *jr4 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
    jr4->SetName("myGraph");
```

```
    jr4->Draw();
  TImage *imj = TImage::Create();
   imj->FromPad(c1);
   imj->WriteImage("canvas2.png");

    TCanvas *c2 = new TCanvas("c2", "General Plots3",800,600);
    c2->Divide(2,2);
    c2->cd(1);
    tree->Draw("pj:pa");
    TGraph *pr = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
    pr->SetName("myGraph");
    pr->Draw();

    c2->cd(2);
    tree->Draw("pk:pa");
    TGraph *pr2 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
    pr2->SetName("myGraph");
    pr2->Draw();

    c2->cd(3);
    tree->Draw("pl:pa");
    TGraph *pr3 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
    pr3->SetName("myGraph");
    pr3->Draw();

    c2->cd(4);
    tree->Draw("pm:pa");
    TGraph *pr4 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
    pr4->SetName("myGraph");
    pr4->Draw();
  TImage *imp = TImage::Create();
   imp->FromPad(c2);
   imp->WriteImage("canvas3.png");
```

```
  TCanvas *c3 = new TCanvas("c3", "General Plots4",800,600);
  c3->Divide(2,2);
  c3->cd(1);
  tree->Draw("pn:pa");
  TGraph *kr = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
  kr->SetName("myGraph");
  kr->Draw();

  c3->cd(2);
  tree->Draw("po:pa");
  TGraph *kr2 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
  kr2->SetName("myGraph");
  kr2->Draw();

  c3->cd(3);
  tree->Draw("pp:pa");
  TGraph *kr3 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
  kr3->SetName("myGraph");
  kr3->Draw();

  c3->cd(4);
  tree->Draw("pq:pa");
  TGraph *kr4 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
  kr4->SetName("myGraph");
  kr4->Draw();
 TImage *imk = TImage::Create();
  imk->FromPad(c3);
  imk->WriteImage("canvas4.png");

  TCanvas *c4 = new TCanvas("c4", "General Plots5",800,600);
  c4->Divide(2,2);
  c4->cd(1);
  tree->Draw("pr:pa");
  TGraph *fr = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
```

```
fr->SetName("myGraph");
fr->Draw();

c4->cd(2);
tree->Draw("ps:pa");
TGraph *fr2 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
fr2->SetName("myGraph");
fr2->Draw();

c4->cd(3);
tree->Draw("pt:pa");
TGraph *fr3 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
fr3->SetName("myGraph");
fr3->Draw();

c4->cd(4);
tree->Draw("pu:pa");
TGraph *fr4 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
fr4->SetName("myGraph");
fr4->Draw();
TImage *imf = TImage::Create();
imf->FromPad(c4);
imf->WriteImage("canvas5.png");

TCanvas *c5 = new TCanvas("c5", "General Plots6",800,600);
c5->Divide(3,2);
c5->cd(1);
tree->Draw("pv:pa");
TGraph *dr = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
dr->SetName("myGraph");
dr->Draw();

c5->cd(2);
tree->Draw("pw:pa");
```

```
    TGraph *dr2 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
    dr2->SetName("myGraph");
    dr2->Draw();

    c5->cd(3);
    tree->Draw("px:pa");
    TGraph *dr3 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
    dr3->SetName("myGraph");
    dr3->Draw();

    c5->cd(4);
    tree->Draw("py:pa");
    TGraph *dr4 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
    dr4->SetName("myGraph");
    dr4->Draw();

    c5->cd(5);
    tree->Draw("pz:pa");
    TGraph *dr5 = new TGraph(tree->GetSelectedRows(),tree->GetV2(), tree->GetV1());
    dr5->SetName("myGraph");
    dr5->Draw();
  TImage *imd = TImage::Create();
   imd->FromPad(c4);
   imd->WriteImage("canvas6.png");
}
```

References

Borne, K. (2014). *Top 10 big data challenges—A serious look at 10 big data V's*. https://mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs/. Accessed January 31, 2018.

Cleverism. (2018). *What is big data?* https://www.cleverism.com/brief-history-big-data/. Accessed January 31, 2018.

Dataversity. (2018). Big Data Trends for 2018, https://www.dataversity.net/big-data-trends-2018/. Accessed November 12, 2018.

Emetere, M. E. (2016a). Statistical examination of the aerosols loading over Mubi-Nigeria: The satellite observation analysis. *Geographica Panonica, 20*(1), 42–50.

Emetere, M. E. (2016b). *Numerical modelling of West Africa regional scale aerosol dispersion*. Thesis submitted to Covenant University.

Emetere, M. E. (2017a). Investigations on aerosols transport over micro- and macro-scale settings of West Africa. *Environmental Engineering Research, 22*(1), 75–86.
Emetere, M. E. (2017b). Lightning as a source of electricity: Atmospheric modeling of electromagnetic fields. *International Journal of Technology, 8*, 508–518.
Emetere, M. E. (2017c). Impacts of recirculation event on aerosol dispersion and rainfall patterns in parts of Nigeria. *Global Nest Journal, 19*(2), 344–352.
Emetere, M. E. (2017d). Monitoring the 3-year thermal signatures of the Calbuco pre-volcano eruption event. *Arabian Journal of Geoscience, 10*, 94. https://doi.org/10.1007/s12517-017-2861-z.
Emetere, M. E., & Akinyemi, M. L. (2017). Documentation of atmospheric constants over Niamey, Niger: A theoretical aid for measuring instruments. *Meteorological Applications, 24*(2), 260–267.
Emetere, M. E., Akinyemi, M. L., & Akinojo, O. (2015a). A novel technique for estimating aerosol optical thickness trends using meteorological parameters. *2015 PIAMSEE: AIP Conference Proceedings, 1705*(1), 020037.
Emetere, M. E., Akinyemi, M. L., & Uno, U. E. (2015b). Computational analysis of aerosol dispersion trends from cement factory. In *IEEE Proceedings 2015 International Conference on Space Science & Communication* (pp. 288–291).
Emetere, M. E., Akinyemi, M. L., & Akinojo, O. (2015c). Parametric retrieval model for estimating aerosol size distribution via the AERONET, LAGOS station. *Environmental Pollution, 207*(C), 381–390.
Emetere, M. E., Akinyemi, M. L., & Akin-Ojo, O. (2015d). Aerosol optical depth pollution in selected areas trends over different regions of Nigeria: Thirteen years analysis. *Modern Applied Science, 9*(9), 267–279.
Emetere, M. E., Akinyemi, M. L., & Edeghe, E. B. (2016). A simple technique for sustaining solar energy production in active convective coastal regions. *International Journal of Photoenergy, 2016*(3567502), 1–11. https://doi.org/10.1155/2016/3567502.
Foote, K. D. (2017). *A brief history of big data*. http://www.dataversity.net/brief-history-big-data/. Accessed January 30, 2018.
Qubole, (2008). The Future of Big Data and Machine Learning Is Clear: It's All on the Cloud, https://www.qubole.com/blog/the-future-of-big-data-and-machine-learning-is-clear-its-all-on-the-cloud/. Accessed November 12, 2018.
Stephenson, D. (2013). 7 big data techniques that create business value. https://www.firmex.com/thedealroom/7-big-data-techniques-that-create-business-value/. Accessed January 31, 2018.